Steffen Hölldobler et al. (Hrsg.)

# Ausgezeichnete Informatikdissertationen 2017

Gesellschaft für Informatik e.V. (GI)

**Lecture Notes in Informatics (LNI) - Proceedings**
Series of the Gesellschaft für Informatik (GI)

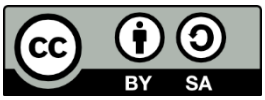
Volume D-18

ISBN 978-3-88579-977-1

**Volume Editors**
Prof. Dr. Steffen Hölldobler
Technische Universität Dresden
01062 Dresden, Deutschland
sh@iccl.tu-dresden.de




**printed by** Köllen Druck+Verlag GmbH, Bonn



# Vorwort

Die Gesellschaft für Informatik e.V. (GI) vergibt gemeinsam mit der Schweizer Informatik Gesellschaft (SI) und der österreichischen Computergesellschaft (OCG) jährlich einen Preis für eine hervorragende Dissertation im Bereich der Informatik. Hierzu zählen nicht nur Arbeiten, die einen Fortschritt in der Informatik bedeuten, sondern auch Arbeiten aus dem Bereich der Anwendungen in anderen Disziplinen und Arbeiten, die die Wechselwirkungen zwischen Informatik und Gesellschaft untersuchen. Die Auswahl dieser Dissertationen stützt sich auf die von den Universitäten und Hochschulen für diesen Preis vorgeschlagenen Dissertationen. Jede dieser Hochschulen kann jedes Jahr nur eine Dissertation vorschlagen. Somit sind die im Auswahlverfahren vorgeschlagenen Kandidatinnen und Kandidaten bereits „Preisträger“ ihrer Hochschule.

Die 28 Einreichungen zum Dissertationspreis 2017 belegen die Bedeutung und auch die Bekanntheit des Dissertationspreises. Wie jedes Jahr wurden die vorgeschlagenen Arbeiten im Rahmen eines Kolloquiums im Leibniz-Zentrum für Informatik Schloss Dagstuhl von den Nominierten vorgestellt. Für die Mitglieder des Nominierungsausschusses war das persönliche Zusammentreffen mit den Nominierten der Höhepunkt der Auswahlarbeit, und für die Nominierten hat das Kolloquium sicher eine Reihe neuer Erfahrungen und wissenschaftlicher Kontakte geboten. Das wissenschaftlich sehr hohe Niveau der Vorträge, die regen Diskussionen und die angenehme Atmosphäre in Schloss Dagstuhl wurde von allen Teilnehmerinnen und Teilnehmern des Kolloquiums sehr begrüßt.

Wie in jedem Jahr fiel es dem Nominierungsausschuss sehr schwer, eine Dissertation auszuwählen, die durch den Preis besonders gewürdigt wird. Mit der Präsentation aller vorgeschlagenen Dissertationen in diesem Band wird die Ungerechtigkeit, eine aus mehreren ebenbürtigen Dissertationen hervorzuheben, etwas ausgeglichen. Dieser Band soll zudem einen Beitrag zum Wissenstransfer innerhalb der Informatik und von den Universitäten und Hochschulen in die Bereiche Technik, Wirtschaft und Gesellschaft leisten.

Die beteiligten Gesellschaften zeichnen Herrn Dr. techn. Daniel Gruss für seine Dissertation „Software-based Microarchitectural Attacks“ und Herrn Dr. Ämin Baumeler für seine Dissertation „Causal Loops: Logically Consistent Correlations, Time Travel, and Computation“ mit dem Dissertationspreis 2017 aus.

Herr Gruss hat eine ingenieur-wissenschaftliche Arbeit im Bereich der Mikroarchitekturangriffe verfasst, in der er zeigt, dass Angriffe vollständig automatisiert werden können, es neue Seitenkanäle gibt und Angriffe auch in stark eingeschränkten Umgebungen und auf jedem Computersystem durchgeführt werden können.

Herr Baumeler hat eine theoretische Arbeit auf höchstem Niveau im Spannungsfeld zwischen Physik und Informatik vorgelegt. Er zeigt, dass in einer Welt, in der quantenmechanische und relativistische Effekte auftreten können, kausale Schleifen bezüglich logischen, physikalischen und berechnungstheoretischen Prinzipien unproblematisch sein können.

Mit dieser Preisverleihung würdigen die beteiligten Gesellschaften – die Gesellschaft für Informatik e.V. (GI), die Schweizer Informatik Gesellschaft (SI) und die österreichische Computergesellschaft (OCG) – zwei herausragende wissenschaftliche Arbeiten. Eine Arbeit mit hoher praktischer Relevanz auf dem Gebiet der Sicherheit moderner Computer-

systeme und eine theoretische Arbeit, die viele unterschiedliche Gebiete miteinander verknüpft, ohne an der Oberfläche zu bleiben und ohne die traditionellen Grenzen der wissenschaftlichen Klassifikation zu beachten.

Ein besonderer Dank gilt dem Nominierungsausschuss, der sehr effizient und konstruktiv zusammengearbeitet hat. Bei Frau Emmanuelle-Anna Dietz Saldanha möchte ich mich für die Unterstützung bei der Entgegennahme der vorgeschlagenen Dissertationen, für die Organisation des Kolloquiums sowie für die Zusammenstellung und Anpassung der Beiträge an das Format der GI-Edition Lecture Notes in Informatik (LNI) bedanken. Für die finanzielle Unterstützung des Nominierungskolloquiums sei den beteiligten Gesellschaften gedankt. Die Gastfreundlichkeit und die hervorragende Bewirtung in Dagstuhl trugen zum Erfolg des Kolloquiums bei, wofür ich mich an dieser Stelle ebenfalls herzlich bedanke.

Steffen Hölldobler
Dresden im August 2018

## Kandidaten für den GI-Dissertationspreis 2017

| | |
|---|---|
| Dr. Baumeler, Ämin | Universitá della Svizzera italiana |
| Dr. Bercher, Pascal | Universität Ulm |
| Dr. techn. Bliem, Bernhard | TU Wien |
| Dr. rer. nat. Borrmann, Dorit | Julius-Maximilians-Universität Würzburg |
| Dr. rer. nat. Ceylan, Ismail Ilkan | Technische Universität Dresden |
| Dr. rer. nat. Gadiraju, Ujwal | Leibniz Universität Hannover |
| Dr. Große, Katharina | Zeppelin Universität |
| Dr. techn. Gruss, Daniel | Technische Universität Graz |
| Dr. Jakobs, Marie-Christine | Universität Paderborn |
| Dr.-Ing. Kaethner, Christian | Universität zu Lübeck |
| Dr.-Ing. Kaltenbrunner, Martin | Bauhaus-Universität Weimar |
| Dr. rer. nat. Keszöcze, Oliver | Universität Bremen |
| Dr. Kurras, Sven | Universität Hamburg |
| Dr. Loitzenbauer, Veronika | Universität Wien |
| Dr.-Ing. Lopacinski, Lukasz | Brandenburgische Technische Universität Cottbus-Senftenberg |
| Dr.-Ing. Matuszyk, Pawel | Otto-von-Guericke-Universität Magdeburg |
| Dr. Müggler, Elias | Universität Zürich |
| Dr.-Ing. Nürnberger, Stefan | Universität des Saarlandes |
| Dr. rer.nat. Papenbrock, Thorsten | Universität Potsdam |
| Dr. rer.nat. Peters, Christoph | Rheinische Friedrich-Wilhelms-Universität Bonn |
| Dr. Pommerening, Florian | Universität Basel |
| Dr. Rädle, Roman | Universität Konstanz |
| Dr. Schwiegelshohn, Chris | TU Dortmund |
| Dr.-Ing. Stab, Christian M. E. | TU Darmstadt |
| Dr.-Ing. Thies, Justus | Friedrich-Alexander Universität Erlangen Nürnberg |
| Dr. Wendler, Philipp | Universität Passau |
| Dr.-Ing. Widmann, Stefan | FernUniversität in Hagen |
| Dr. rer. nat. Wieczorek, Matthias | Technische Universität München |

## Mitglieder des Nominierungsausschusses für den GI-Dissertationspreis 2017

Von links nach rechts:

| | |
|---|---|
| Prof. Dr. Gustaf Neumann | Wirtschaftsuniversität Wien |
| Prof. Dr.-Ing. Wolfgang Effelsberg | Universität Mannheim |
| Prof. Dr. Paul Molitor | Martin-Luther-Universität Halle-Wittenberg |
| Prof. Dr. Sven Apel | Universität Passau |
| Prof. Dr.-Ing. Felix Freiling | Universität Erlangen-Nürnberg |
| Prof. Dr. Myra Spiliopoulou | Otto-von-Guericke-Universität Magdeburg |
| Prof. Dr. Nicole Schweikardt | Humbold-Universität zu Berlin |
| Prof. Dr. Abraham Bernstein | Universität Zürich |
| Prof. Dr. Björn Scheuermann | Humbold-Universität zu Berlin |
| Prof. Dr. Steffen Hölldobler (Vorsitzender) | Technische Universität Dresden |

Nicht im Bild:

| | |
|---|---|
| Prof. Dr. Hans-Peter Lenhof | Universität des Saarlandes |
| Prof. Dr. Rüdiger Reischuk | Universität zu Lübeck |
| Prof. Dr. Sabine Süsstrunk | École Polytechnique Fédérale de Lausanne |

# Inhaltsverzeichnis

# Causal Loops: Logically Consistent Correlations, Time Travel, and Computation

Ämin Baumeler[1]

## 1 Einleitung, Motivation und Resultate

Kausale Schleifen sind Schleifen in Ursache-Wirkung Beziehungen, wobei eine Wirkung die Ursache der Ursache ebendieser Wirkung ist. Diese Dissertation handelt von solchen kausalen Schleifen mit der Aussage, dass sie unproblematisch (im logischen und im berechnungstheoretischen Sinn) sein können.

Die gängigen physikalischen Theorien, also die Quantentheorie, die allgemeine Relativitätstheorie und eine mögliche Theorie der Quantengravitation, motivieren dieses Studium. Im starken Gegensatz zur klassischen Physik (Newton-Mechanik oder Relativitätstheorie) erlaubt die Quantentheorie Superpositionen von Zuständen. Das übliche Beispiel um eine Superposition zu veranschaulichen ist die Überlagerung zweier Wellenfunktionen desselben quantenmechanischen Systems wo der Aufenthaltsort in den jeweiligen Wellenfunktionen unterschiedlich ist; das System befindet sich in einer Superposition in Bezug auf den Ort. Man könnte sich vorstellen dieses Superpositionsprinzip auf andere Freiheitsgrade zu erweitern; im speziellen auf die zeitliche Anordnung quantenmechanischer Systeme. Dies würde zu einer Superposition unterschiedlicher kausalen Strukturen führen. Im Gedankenexperiment könnte man könnte man einen Planeten (als Konglomerat quantenmechanischer Systeme betrachtet) in eine Superposition bezüglich dessen Position setzen, wobei die relativistische Raumzeit so verformt wird, dass eine Superpositionen unterschiedlicher kausalen Strukturen entsteht.

Wir erläutern eine weitere der Quantentheorie entspringenden Motivation dieses Studiums. Die Quantentheorie widerspricht unserer Intuition und beruht auf Postulaten mathematischer Natur, die sich bisweilen physikalischen Interpretation entziehen. In diesem Zusammenhang stellt sich die Frage, ob ein anderer Zugang zur Theorie, wo die Kausalität abgeschwächt wird, zu einer Vereinfachung führen würde; ein solcher Ansatz wurde bisweilen nicht verfolgt, scheint jedoch natürlich (die Quantentheorie behandelt die Zeit auf eine grundsätzlich unterschiedliche Weise als andere physikalische Grössen).

Es ist bekannt, dass die allgemeine Relativitätstheorie zyklische Raumzeitgeometrien zulässt. Die Frage nach solchen Geometrien wurde ursprünglich von Einstein aufgestellt und deren Existenz innerhalb der Theorie von Lanczos [La24], Gödel [Gö49] und weiteren WissenschaftlerInnen erwiesen. Weltlinien, *d.h.*, Trajektorien in einer solchen Geometrie können

[1] IQOQI Vienna, Boltzmanngasse 3, 1090 Vienna, Austria, aemin.baumeler@univie.ac.at

zeitlich geschlossen sein: Ein Teilchen auf einer solchen Bahn würde also *mit sich selbst* zusammenstossen. Die Frage, die sich hier stellt, ist nun, ob diese Lösungen der Relativitätstheorie als mathematische Artefakte angesehen werden können und ob es logische, physikalische oder berechnungstheoretische Gründe gibt, Zeitreisen dieser Art zu verwerfen.

Diese Überlegungen habe wir in drei Teilbereichen weiter verfolgt: in Bezug auf *Korrelationen*, *Zeitreisen* und *Berechnungen*. Wir konnten zeigen, dass *logisch-konsistente Korrelationen* nicht zwingend kausal sein müssen. Geometrisch bilden diese Korrelationen ein Polytop, welches wir charakterisieren konnten [BW16b]. Zusätzlich konnten wir diese Korrelationen als Lösungen eines Fixpunkt-Problems darstellen [BW16a]. Diese Arbeit zeigt uns, dass es eine Welt *ausserhalb des Kausalen* und *innerhalb des Logischen* gibt. Dies führte uns zur erweiterten Analyse von Zeitreisen. Die vorhergehenden Analysen [Fr90, EKT91] ignorierten die freie Wahl der Zeitreisenden. Diese Freiheit haben wir mathematisch modelliert und konnten dadurch zeigen, dass Lösungen für Zeitreisen dadurch *physikalisch* unproblematisch werden [Ba17]. Letztendlich hatten wir uns gefragt, welche Probleme eine zeitreisende Rechenmaschine effizient lösen kann. Wäre die Rechnungsleistung zu stark (gäbe es keine Trennung zwischen den entsprechenden Komplexitätsklassen von $\mathsf{P}$ und $\mathsf{NP}$), so könnte dies als Argument *gegen* Akausalität verstanden werden [Aa08]. Wir konnten aber zeigen, dass die Rechenleistung solcher Maschinen durch eine Komplexitätsklasse zwischen $\mathsf{P}$ und $\mathsf{NP}$ beschränkt ist: Sie ist nicht stärker als $\mathsf{UP} \cap \mathsf{coUP}$ [BW18]. Dieses Resultat erlaubt uns zudem die bereits bekannte Klasse $\mathsf{UP} \cap \mathsf{coUP}$ alternativ zu charakterisieren. Das Fazit dieser Überlegungen ist, dass Akausalität aus *logischer*, *physikalischer* und *berechnungstheoretischer* Sicht unproblematisch ist.

## 2 Zur Kausalität, Annahmen und mögliche Widersprüche

In der Physik und in der Philosophie der Physik wird der Begriff der Kausalität oft debattiert [BHM12]. Die Dissertation gibt keine detaillierte Beschreibung des Begriffs wieder, sondern konzentriert sich auf *Ursache-Wirkung* Relationen. In dieser Dissertation verstehen wir den Begriff der Kausalität als eine Relation zwischen Zufallsvariablen. Ein solcher Begriff der Kausalität wird oft mit dem Begriff der *freien Wahl* in Verbindung gesetzt: Eine Zufallsvariable $A$ ist *frei* falls sie unabhängig von allen Variablen *ausserhalb des Zukunft-Lichtkegels* von $A$ ist [CR11]. Die Umkehrung dieser Definition von freien Variablen ermöglicht eine *Definition der Kausalität*, was wir im Artikel [BW14] weiter verfolgt haben.

**Definition 1** (Kausale Zukunft/Vergangenheit, Ursache/Wirkung)**.** Eine Zufallsvariable $A$ ($B$) liegt in der *kausalen Zukunft (Vergangenheit)* von $B$ ($A$) dann und nur dann, wenn beide Variablen korreliert sind und $A$ *frei* ist. Diese Relation wird durch $A \preceq B$ ausgedrückt. Die Variable $A$ nennen wir in diesem Fall eine *Ursache* der Variable $B$, der *Wirkung*.

Hier ist zu beachten, dass die Definition auf einem Begriff der *Freiheit* aufbaut, welcher nicht weiter erläutert wird.[2] Diese Definition ist in der Abbildung 1 illustriert: Ist der Zustand eines Drehknopfs mit der Position eines Zeigers korreliert und kann der Drehknopf frei eingestellt werden, so ist diese Einstellung (Zustand des Drehknopfs) in der kausalen Vergangenheit des Zustands des Zeigers. Eine ähnliche Definition lässt sich auch in der

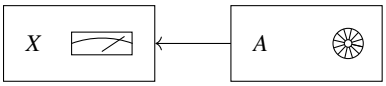



Abb. 1: Falls $A$ und $X$ korreliert sind und $A$ frei ist, ist $A$ in der kausalen Vergangenheit von $X$.

Philosophie unter dem Namen *Interventionsit's appraoch to causality* auffinden [BHM12].

Eine häufige, oft implizite, Annahme bezüglich der Kausalität ist, dass kausale Relationen eine *definite partielle Ordnung* von Ereignissen oder Zufallsvariablen widerspiegeln. Die Eigenschaft der *partiellen Ordnung* besagt, dass keine Zyklen auftreten; die der *definiten Ordnung*, dass die Relationen vorherbestimmt sind und im speziellen sich nicht in einer Art Superposition befinden. Die obige Definition der kausalen Beziehungen, hingegen, erlaubt uns diese Annahme zu umgehen: Zyklen sind nicht ausgeschlossen und der Wert der Zufallsvariablen könnte von Quantenmechanischen Systemen abhängen. Verwerfen wir diese eben besprochene Annahme, so stellt sich die unverzügliche Frage nach Paradoxien: *Sind kausale Schleifen logisch konsistent?* Wir erläutern zwei Hauptprobleme von solchen Schleifen: das *Grossvater-Paradox* und das *Informations-Paradox*.

Das *Grossvater-Paradox* entsteht, wenn eine Wirkung ihre eigene Ursache auslöscht. Könnte man in die Vergangenheit reisen (was man sich als kausale Schleife vorstellen kann), so könnte man den eigenen Grossvater vor der eigenen Geburt ermorden; dadurch kann man aber nicht geboren werden und daher auch nicht in die Vergangenheit reisen: Wir haben es hier mit einem logischen Widerspruch zu tun.

Das *Information-Paradox* ist die Kehrseite des Grossvater-Paradoxes: Mehrere potentielle Wirkungen bestätigen ihre eigene Ursache, ohne dass ein Ursache-Wirkung Paar ausgesondert wird. Hierzu kann man die Geschichte einer Zeitreisenden erzählen, die an einem Morgen ein Buch neben ihrem Bett auffindet, das Buch veröffentlicht, in die Vergangenheit reist um das Buch neben ihr eigenes Bett abzulegen. Wer hat das Buch geschrieben und was ist der Inhalt?

Physikalisch gesehen entstehen dieses Paradoxien, falls eine Theorie, gegeben der Anfangsbedingungen und der Dynamik, keine Voraussagen treffen kann, auch keine probabilistische. Im ersten Fall versagt die Aussagekraft einer solchen Theorie, da es *keine* konsistenten Lösungen gibt; im zweiten, da es *mehrere* konsistente Lösungen gibt. Im folgenden werden wir uns vor diesen Paradoxien in acht nehmen.

[2] Eine Möglichkeit den Begriff der Freiheit zu definieren haben wir im Artikel [BW17a] verfolgt.

## 3 Kausale Korrelationen und kausale Ungleichungen

Der oben definierte Begriff der Kausalität wird nun mit einem Begriff von Parteien zusammengebracht. Man kann sich eine Partei als eine Person vorstellen, die in einem geschlossenen Rahmen Operationen durchführen kann.

**Definition 2** (Partei)**.** Eine *Partei* $S_j = (A_j, X_j, L_j)$ ist ein Tripel bestehend aus einer *freien* Zufallsvariablen $A_j$, einer nicht freien Zufallsvariablen $X_j$ und einer lokalen Operation $L_j$. Die freie Variable nennen wir *Eingabe* und die andere *Ausgabe*. Sind $S_j$ und $S_k$ Parteien, dann definieren wir die Relation $S_j \preceq S_k$ als äquivalent zu $A_j \preceq X_k$.

Wir erweitern unsere Sprache auf *kausale Korrelationen*. Wir nennen Korrelationen kausal, falls diese durch eine *definite partielle Ordnung* der Parteien erzeugt werden können. In diesem Zusammenhang studieren wir die Verteilungen $P_{X_1,X_2,\ldots|A_1,A_2,\ldots}$ wobei die Variablen $X_j$ und $A_j$ der Partei $S_j$ angehören.

**Definition 3** (Kausale Korrelationen, akausale Korrelationen)**.** Im Fall von zwei Parteien nennen wir eine Verteilung $P_{X_1,X_2|A_1,A_2}$ *kausal* dann und nur dann, wenn sie sich als

$$P_{X_1,X_2|A_1,A_2} = pP_{X_1|A_1}P_{X_2|X_1,A_1,A_2} + (1-p)P_{X_1|X_2,A_1,A_2}P_{X_2|A_2}\,, \quad 0 \leq p \leq 1\,,$$

zerlegen lässt. Diese Zerlegung besagt, dass die Partei $S_1$ mit Wahrscheinlichkeit $p$ in der Vergangenheit von $S_2$ liegt. Kann eine Zwei-Parteien-Verteilung nicht auf diese Art geschrieben werden, so ist sie *akausal*.

Eine *kausale Ungleichung* ist ein Ausdruck, mit dem wir überprüfen können, ob eine solche Zerlegung möglich ist. Als Beispiel betrachten wir die Wahrscheinlichkeit, dass jede Partei die Eingabe der anderen Partei ausgibt: $\Pr(X_1 = A_2 \wedge X_2 = A_1)$. Die maximale Grösse dieses Ausdrucks für kausale Verteilungen und binäre Zufallsvariablen mit uniform verteilten Eingaben ist $1/2$: Ist $S_1$ in der Vergangenheit von $S_2$, so ist $\Pr(X_1 = A_2) = 1/2$ und $\Pr(X_2 = A_1)$ beliebig ($S_1$ könnte ihre Eingabe an $S_2$ weitergeben). Bei der entgegengesetzten Anordnung ist der Ausdruck auch maximal $1/2$, somit auch bei jeder konvexen Mischung beider Anordnungen. Also haben wir es hier mit der kausalen Ungleichung $\Pr(X_1 = A_2 \wedge X_2 = A_1) \leq 1/2$ zu tun.

Im Jahr 2012 hatten Oreshkov, Costa und Brukner [OCB12] gezeigt, dass eine Quantentheorie, die nur auf *lokalen* Annahmen beruht, solche Ungleichungen verletzen kann. Zusätzlich hatten sie gezeigt, dass im klassischen Limit, *d.h.* wenn die lokalen Operationen stochastische Operationen auf klassischen Zufallsvariablen sind, im Zwei-Parteien-Fall nur kausale Korrelationen auftreten. Das hatte die Autoren zur Vermutung gebracht, dass unsere Erfahrung von kausalen Abläufen aus dem Übergang von Quanten- zu klassischen Systemen hervortritt. Wir haben aber gezeigt, dass dies nicht der Fall ist [BFW14]: Im Mehr-Parteien-Fall sind Verletzungen von kausalen Ungleichungen auch mit klassischen Systemen möglich. Dazu präsentieren wir zuerst das Modell, auf dessen Basis die Korrelationen zu Stand kommen. Die Annahmen dieses Modells sind *lokaler* Natur. Die erste Annahme ist, dass Parteien isoliert sind, *d.h.* sie können nicht direkt miteinander

kommunizieren. Die zweite, dass Parteien Operationen auf klassischen Zufallsvariablen durchführen (anstelle von Quantenzuständen). Die dritte, dass jede Partei nur eine Operation durchführt, also nicht an mehreren Raum-Zeit-Punkten mit der Umgebung interagiert. Die letzte, dass die Beobachtungen der Parteien als Wahrscheinlichkeitsverteilung beschrieben werden kann. Eine Partei $S_j$ transformiert dementsprechend eine Zufallsvariable aus der Umgebung ($I_j$) in Abhängigkeit ihrer Eingabe ($A_j$) und generiert damit eine Zufallsvariable $O_j$ die an die Umgebung abgegeben wird und eine Ausgabe $X_j$. Die lokale Operation $L_j$ dieser Partei ist also eine bedingte Verteilung $P_{X_j,O_j|A_j,I_j}$. Daraus leiten wir die allgemeinste Funktion ab, welche die Wahl der lokalen Operationen $L_1, L_2, \ldots$ auf eine Verteilung $P_{X_1,X_2,\ldots|A_1,A_2,\ldots}$ abbildet. Wie wir gezeigt haben, kann eine solche Funktion als Umgebung verstanden werden, welche die $O_1, O_2, \ldots$ Variablen in $I_1, I_2, \ldots$ Variablen transformiert: also als Verteilung $P_{I_1,I_2,\ldots|O_1,O_2,\ldots}$ (siehe Abbildung 2a). Die Umgebungen

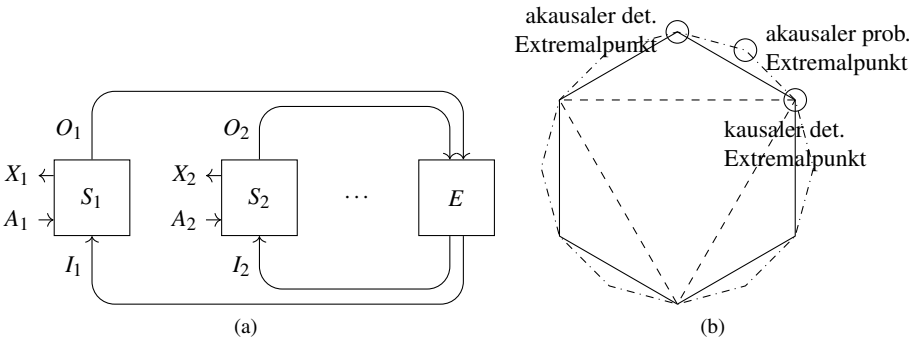



Abb. 2: (a) Die Umgebung ist ein Kanal welcher die $O$ Variablen in $I$ Variablen transformiert. (b) Das mittige Dreieck repräsentiert alle Umgebungen, die zu kausalen Verteilungen führen. Das äusserste Polytop enthält alle logisch-konsistenten Umgebungen: Das Polytop ist grösser als das Dreieck. Das Polytop mit den durchgezogenen Linien ist das Polytop der logisch-konsistenten Umgebungen, wo die Extremalpunkte deterministische Transformationen beschreiben.

dieser Art bilden ein Polytop [BW16b], welches wir schematisch in der Abbildung 2b zeigen.

In dieser Kurzfassung besprechen wir einen akausalen Extremalpunkt dieses Polytops, der probabilistisch ist. Das Beispielsszenario besteht aus drei Parteien $(S_1, S_2, S_3)$ und binären Zufallsvariablen, die zwischen den Parteien und der Umgebung ausgetauscht werden. Die Parteien, die wir beachten, haben jeweils eine binäre Ausgabe und eine Paar von Eingaben: eine binäre Eingabe $A_j$ und eine gemeinsame Eingabe $M$, welche die Werte $1,2,3$ annehmen kann; die Eingaben sind uniform verteilt. Eine kausale Ungleichung für dieses Szenario ist

$$\frac{1}{3}\left(\Pr(X_1 = A_2 \oplus A_3 | M = 1) + \Pr(X_2 = A_1 \oplus A_3 | M = 2) + \Pr(X_3 = A_1 \oplus A_2 | M = 3)\right) \leq \frac{5}{6}.$$

Der Maximalwert dieses Ausdruckes ist $5/6$, da in einer kausalen Konfiguration mindestens eine Partei keine anderen Parteien in ihrer Vergangenheit hat: mindestens einer der

drei Wahrscheinlichkeitausdrücke ist $1/2$. Nehmen wir hingegen die Umgebung, welche in Abbildung 3 dargestellt ist, existiert eine Wahl von lokalen Operationen $L_1, L_2, L_3$ wo der Ausdruck den Wert 1 erreicht. Diese Umgebung beschreibt mit halber Wahrscheinlich-

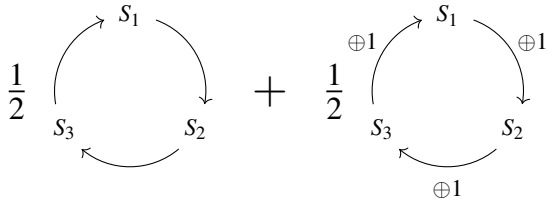


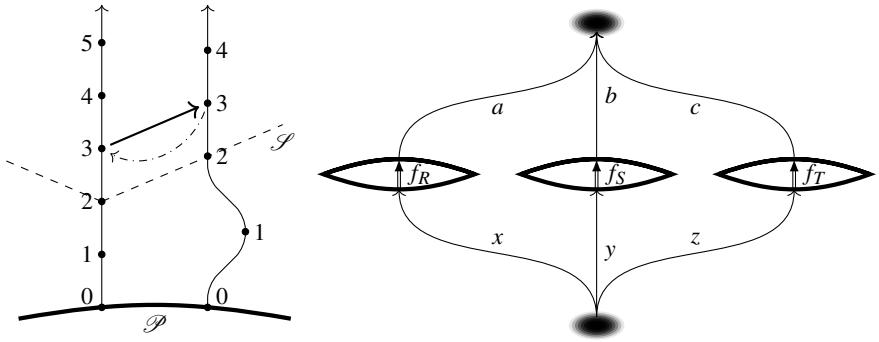

Abb. 3: Umgebung, mit welcher wir die genannte Drei-Parteien-Ungleichung maximal verletzen können.

keit einen zyklischen Identitätskanal: Die Variable $I_2$ hat denselben Wert wie $O_1$ und so weiter. Mit halber Wahrscheinlichkeit ist der zyklische Kanal negierend: Die Variable $I_2$ hat den negierten Wert wie $O_1$ und so weiter. Jede Wahl von lokalen Operationen der Parteien resultiert unter Verwendung dieser Umgebung in einer gültigen Verteilung. Falls die Parteien nun die folgenden lokalen Operationen verwenden, so wird die oben beschrieben Ungleichung maximal verletzt. Hat die Eingabe $M$ den Wert 1, so gibt Partei $S_1$ den Wert 0 an die Umgebung, Partei $S_2$ gibt den Wert $A_2$ an die Umgebung und Partei $S_3$ addiert $A_3$ zu $I_3$ (wobei die Addition modulo 2 durchgeführt wird) und sendet diese Summe an die Umgebung. Bei dieser Wahl von lokalen Operation ist $I_1 = A_2 \oplus A_3$. In den Fällen, wo $M$ einen andern Wert annimmt rotieren die Partien die lokale Operation dementsprechend.

## 4 Zeitreisen

Forscher um Thorne [EKT91] und Novikov [Fr90] hatten sich überlegt, wie eine Billardkugel sich verhält, wenn sie mit sich selbst Zusammenstosst. Dazu hatten sie sich Geometrien der Relativitätstheorie angeschaut, die Zeitreisen erlauben. Eine Weltlinie für Zeitreisen wird in ihrem Modell wie folgt hergestellt: Man nehme zuerst ein Wurmloch (welche auch in der Relativitätstheorie auftreten) und beschleunigt das eine Ende dieses Wurmlochs auf eine hohe Geschwindigkeit und wieder zurück auf den ursprünglichen Ort. Durch die Zeitdilatation entstehen, wie in Abbildung 4a dargestellt, zeitlich geschlossene Weltlinien. Eine Billardkugel auf einer solchen Weltlinie kollidiert mit sich selbst. Ihr Ergebnis ist, dass für jede Anfangsbedingung der Billardkugel konsistente Trajektorien existieren, oft eine *unendliche* Anzahl von Trajektorien. Die erinnert uns an das oben besprochene Informations-Paradox. Wir haben uns mit derselben Geometrie beschäftigt und gefragt, was passiert wenn ExperimentatorInnen zusätzlich zur Spezifikation der Anfangsbedingungen die Bahn der Kugel beliebig ändern können [Ba17]. Wie verändert die Wahl der Manipulation der Bahn die Anzahl an konsistenten Lösungen?

Unser Modell baut auf zwei Prinzipien: auf *Novikovs Prinzip* [Fr90] der Selbst-Konsistenz und auf dem Prinzip *"keine neue Physik"*. Novikovs Prinzip besagt, dass nur konsistente Trajektorien auftreten. Führt eine Weltline zu einem logischen Widerspruch im Sinn vom Grossvater-Paradox, so ist diese Weltline ausgeschlossen; keine Lösung der Gleichun-




(a) (b)

Abb. 4: (a) Ein Wurmloch erzeugt eine Geometrie mit zeitlich geschlossenen Weltlinien in der Zukunft der Fläche $\mathscr{S}$. (b) Die schwarzen Ovale beschrieben den Eintritts- und Austrittspunkt einer Zeitreise. In den lokalen Regionen $R,S,T$ können beliebige Funktionen durchgeführt werden.

gen. Das zweite Prinzip besagt, dass die physikalischen Gesetze in lokalen Raum-Zeit-Regionen unabhängig von möglichen zeitlich geschlossenen Weltlinien sind. Die Forscher um Novikov und Thorne hatten auch beide Prinzipien beachtet, wobei das letztere jedoch auf die Anfangsbedingungen beschränkt ist: Die physikalischen Gesetze in einer lokalen Region *bevor* zeitlich geschlossenen Weltlinien entstehen sind dieselben wie in einer Welt wo Zeitreisen unmöglich sind. Das bedeutet, dass ExperimentatorInnen die Anfangsbedingung der Billardkugel beliebig wählen können; sie sind in dieser Wahl uneingeschränkt; zu *jeder* Wahl der Anfangsbedingungen existieren konsistente Bahnen.

Wir haben das zweite Prinzip auf lokale Regionen auf den Weltlinien erweitert [Ba17]: ExperimentatorInnen können die Anfangsbedingungen und *zusätzlich* den Zustand der Billardkugel auf ihrer Weltlinie beliebig verändern. Um diese Frage anzugehen, beschreiben wir eine Anzahl von lokalen Raum-Zeit-Regionen, wo beliebige Funktionen durchgeführt werden können. Jede solche Region können wir in eine Vergangenheits- und Zukunftsfläche aufspalten (siehe Abbildung 4b). Jede Vergangenheitsfläche und Zukunftsfläche hat einen Zustandsraum. In dieser Diskussion betrachten wir das Beispiel mit drei lokalen Regionen. Die Zustandsräume sind $\mathscr{X},\mathscr{A}$ für die lokale Region $R$, und so weiter. Somit ist die Funktion $f_R$ eine Funktion von $\mathscr{X}$ auf $\mathscr{A}$. In diesem Beispiel ist die Zeitreise eine Funktion $e : \mathscr{A}\times\mathscr{B}\times\mathscr{C} \to \mathscr{X}\times\mathscr{Y}\times\mathscr{Z}$. In Übereinstimmung zu beiden Prinzipien verlangen wir nach einer Existenz eines Fixpunktes $(i_1,i_2,i_3)$ für jede Wahl der lokalen Operationen: $\forall f_R,f_S,f_T \exists i_1,i_2,i_3 : (i_1,i_2,i_3) = e\circ(f_R(i_1),f_S(i_2),f_T(i_3))$. Daraus folgt, wie wir zeigen konnten, dass der Fixpunkt *eindeutig* ist: Das Informations-Paradox tritt *nicht* auf.

Eine Lösung der obigen Gleichung für reelle Zustandsräume ist

$$e : \mathbb{R}^3 \to \mathbb{R}^3 , \quad (a,b,c) \mapsto (\Theta(-b)\Theta(c), \Theta(-c)\Theta(a), \Theta(-a)\Theta(b)) ,$$
$$\Theta(r) = \begin{cases} 1 & \text{wenn } r > 0, , \\ 0 & \text{sonst.} \end{cases}$$

Das Vorzeichen der reellen Grösse auf einer Zukunftsfläche bestimmt somit die Signalrichtung zwischen den anderen zwei Regionen. Ist $a$ positiv, so haben wir $y = \Theta(-c)$ und $z = 0$, und so weiter. Hier haben wir es somit mit einer kausalen Schleife zu tun: Die Werte $x, y, z$ hangen von den zukünfitigen Werten $a, b, c$ ab. Zudem entstehen, wie oben beschrieben, unter jeder Wahl der Funkionen keine Paradoxien auf. Zusätzlich, was wir in dieser Kurzfassung nicht näher erleutern, können wir jede solche Zeitreise als *reversible* Transformation beschreiben, womit die Trajektorien auch mit den Prinzipien der Thermodynamik in einklang sind.

## 5 Akausale Berechnung

Man kann sich nun fragen, wie stark die Rechenleistung solcher kausalen Schleifen sind. Dazu beschreiben wir ein akausales Computermodell [BW17b]. Dieses Modell beschreibt Berechnungen mit einem Schaltkreis, wobei die Gatter fixiert sind und die Verbindungen zwischen den Gatter akausal sein können. Zur Illustrationen betrachten Sie die Abbildung 5. Wir nennen einen Schaltkreis logisch konsistent falls ein *eindeutiger Fixpunkt* auf

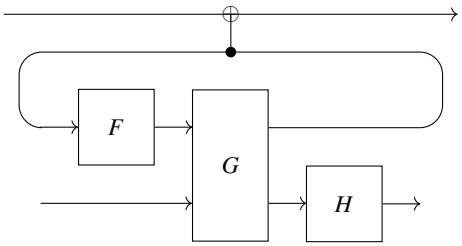



Abb. 5: Akausaler Schaltkreis. Die Verbindung vom Gatter $G$ zum Gatter $F$ ist als eine Verbindung in die Vergangenheit zu verstehen.

den akausalen Verbindung existiert. Dies ist in Einklang mit den akausalen Korrelationen und den Zeitreisen, wo logische Konsistenz mit der Existenz eines eindeutigen Fixpunktes identifiziert werden kann [BW16a]. Im Unterschied zu den oben besprochenen Modellen sind hier jedoch die Funktionen, also die Gatter, fixiert. Wir haben es nicht mehr mit ExperimentatorInnen zu tun, die die Operation frei wählen können. Dadurch ist die Rechenleistung der besprochenen Modellen höchstens so stark wie dieses. In unserer Arbeit konnten wir die Komplexitätsklasse dieses Modell exakt beschreiben [BW18]: $\mathsf{UP} \cap \mathsf{coUP}$. Diese Komplexitätsklasse enthält unter anderem das Faktorisierungsproblem: Gibt es einen Primfaktor kleiner als $k$ einer Zahl $N$? Somit kann dieses akausale Computermodell effizient Zahlen in ihre Primfaktoren zerlegen. Die Klasse $\mathsf{UP} \cap \mathsf{coUP}$ bettet sich wie folgt in die Landschaft der Komplexitätsklassen: $\mathsf{P} \subseteq \mathsf{BPP} \subseteq \mathsf{UP} \cap \mathsf{coUP} \subseteq \mathsf{NP} \cap \mathsf{coNP} \subseteq \mathsf{NP} \subseteq \mathsf{PostBQP} \subseteq \mathsf{PSPACE}$.

Frühere Computermodelle mittels Zeitreisen können, im Gegensatz zu unserem Modell, NP-vollständige Probleme effizient lösen [AW09, Ll11]. Dies macht diese alternativen Modelle zum einen *unphysikalisch* da sie gegen die Annahme der Existenz von schwierig zu lösenden und einfach zu verifizierenden Problemen verstösst [Aa08], zum anderen tritt bei diesen Alternativen das Informations-Paradox auf.

## 6 Schlussfolgerungen

Kausale Schleifen können unproblematisch bezüglich logischen, physikalischen und berechungstheoretischen Prinzipien sein. Dieses Ergebnis zeigt nicht nur, dass Zeitreisen unproblematischer sind als bisher angenommen sondern, öffnet auch das Feld, um selbstbezügliche Systeme zu studieren. Selbstbezüglichkeiten sind grösstenteils in unterschiedlichen Forschungsgebieten *a priori* ausgeschlossen. Als Beispiel können wir hier Tarskis Sprachtheorie [Ta56] nennen. In dieser Theorie sind Sätze wie "Dieser Satz ist falsch" axiomatisch ausgeschlossen. Dieser als Lügner-Paradox bekannte Satz ist problematisch: Ist der Satz wahr, so ist er falsch, ist er falsch, so ist er wahr. Wir haben aber gezeigt, dass Selbstbezüglichkeit nicht zwingend problematisch ist; es gibt eine selbstbezügliche Welt die paradoxfrei ist. In Bezug auf Sprache, können wir zwei sich referenzierende Sätze aufstellen, wo nur eine Wahrheitswertzuweisung gültig ist, und somit weder das Grossvater- noch das Informations-Paradox auftreten:

$R_1$ :"Die Wahrheitswerte dieses und des folgenden Satzes sind unterschiedlich."

$R_2$ :"Die Wahrheitswerte des vorangehenden und dieses Satzes sind gleich."

Die einzige gültige Wahrheitswertzuweisung ist, dass $R_1$ wahr ist und $R_2$ falsch.

## Literaturverzeichnis

[Aa08] Aaronson, Scott: The Limits of Quantum Computers. Scientific American, 298(3):62–69, Marz 2008.

[AW09] Aaronson, Scott; Watrous, John: Closed timelike curves make quantum and classical computing equivalent. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 465(2102):631–647, Februar 2009.

[Ba17] Baumeler, Ämin; Costa, Fabio; Ralph, Timothy C; Wolf, Stefan; Zych, Magdalena: Reversible time travel with freedom of choice. preprint arXiv:1703.00779 [gr-qc], Marz 2017.

[BFW14] Baumeler, Ämin; Feix, Adrien; Wolf, Stefan: Maximal incompatibility of locally classical behavior and global causal order in multiparty scenarios. Physical Review A, 90(4):042106, Oktober 2014.

[BHM12] Beebee, Helen; Hitchcock, Christopher; Menzies, Peter, Hrsg. The Oxford Handbook of Causation. Oxford University Press, Oxford, 2012.

[BW14] Baumeler, Ämin; Wolf, Stefan: Perfect signaling among three parties violating predefined causal order. In: 2014 IEEE International Symposium on Information Theory. IEEE, Piscataway, S. 526–530, Juni 2014.

[BW16a] Baumeler, Ämin; Wolf, Stefan: Device-independent test of causal order and relations to fixed-points. New Journal of Physics, 18(3):035014, April 2016.

[BW16b] Baumeler, Ämin; Wolf, Stefan: The space of logically consistent classical processes without causal order. New Journal of Physics, 18(1):013036, Januar 2016.

[BW17a] Baumeler, Ämin; Wolf, Stefan: Causality-Complexity-Consistency: Can Space-Time Be Based on Logic and Computation? In (Renner, Renato; Stupar, Sandra, Hrsg.): Time in Physics, S. 69–101. Birkhäuser, Cham, 2017.

[BW17b] Baumeler, Ämin; Wolf, Stefan: Non-Causal Computation. Entropy, 19(7):326, Juli 2017.

[BW18] Baumeler, Ämin; Wolf, Stefan: Computational tameness of classical non-causal models. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science, 474(2209):20170698, Januar 2018.

[CR11] Colbeck, Roger; Renner, Renato: No extension of quantum theory can have improved predictive power. Nature Communications, 2:411, August 2011.

[EKT91] Echeverria, Fernando; Klinkhammer, Gunnar; Thorne, Kip S: Billiard balls in wormhole spacetimes with closed timelike curves: classical theory. Physical Review D, 44(4):1077–1099, August 1991.

[Fr90] Friedman, John; Morris, Michael S; Novikov, Igor Dmitriyevich; Echeverria, Fernando; Klinkhammer, Gunnar; Thorne, Kip S; Yurtsever, Ulvi: Cauchy problem in spacetimes with closed timelike curves. Physical Review D, 42(6):1915–1930, September 1990.

[Gö49] Gödel, Kurt: An example of a new type of cosmological solutions of Einstein's field equations of gravitation. Reviews of Modern Physics, 21(3):447–450, Juli 1949.

[La24] Lanczos, Kornel: Über eine stationäre Kosmologie im Sinne der Einsteinschen Gravitationstheorie. Zeitschrift für Physik, 21(1):73–110, Dezember 1924.

[Ll11] Lloyd, Seth; Maccone, Lorenzo; Garcia-Patron, Raul; Giovannetti, Vittorio; Shikano, Yutaka: Quantum mechanics of time travel through post-selected teleportation. Physical Review D, 84(2):025007, Juli 2011.

[OCB12] Oreshkov, Ognyan; Costa, Fabio; Brukner, Časlav: Quantum correlations with no causal order. Nature Communications, 3:1092, Oktober 2012.

[Ta56] Tarski, Alfred: Logic, Semantics, Metamathematics. Clarendon Press, Oxford, 1956.

**Ämin Baumeler** wurde am 4. Juni 1987 in Bern, Schweiz, geboren. Er studierte Informatik mit dem Schwerpunkt auf theoretischer Informatik an der ETH-Zürich. Seine Masterarbeit hatte er am Institute for Quantum Computing der Univerisität Waterloo, Waterloo (ON), Kanada, im Gebiet der Quantenkryptographie geschrieben. Nach dem Masterabschluss, im Jahr 2012, ist er dem Forschungsteam von Prof. Stefan Wolf an der Università della Svizzera italiana, Lugano, Schweiz beigetreten, um da zu promovieren. Die Kerngebiete der Dissertation waren der Begriff der Kausalität sowie dessen Abschwächungen und die Einflüsse einer Abschwächung der Kausalität auf die Physik und Informatik. Im Jahr 2017 hatte er seine Dissertation erfolgreich verteidigt und arbeitet seither als Post-Doc am Institut für Quantenoptik und Quateninformation in Wien.

# Hybrides Planen — Von der Theorie zur Praxis[1]

Pascal Bercher[2]

**Abstract:** Die Dissertation legt Grundlagen, die es erlauben, Planungstechnologie der Künstlichen Intelligenz als Basis für flexible Assistenzsysteme einzusetzen. Die Aufgabe der automatischen Handlungsplanung ist es hierbei, selbständig einen Plan zu entwickeln, der dem Nutzer Schritt für Schritt präsentiert wird und ihn oder sie bei der Bearbeitung einer entsprechenden Aufgabe anleitet. Durch die starke Miteinbeziehung eines menschlichen Nutzers ergeben sich viele neue Herausforderungen: Pläne müssen schnell gefunden werden; und sie sollen nicht nur korrekt sein, sondern auch kostengünstig und dem Nutzer plausibel erscheinen; und sie sollen erklärbar sein, um Transparenz zu schaffen. Aus diesem Grund wurde das hybride Planen gewählt, ein hierarchischer, nicht-linearer Planungsansatz. Es wurden neue Komplexitätsergebnisse für das Planexistenz- und das Planverifikationsproblem erzielt; die ersten zulässigen Heuristiken erforscht, welche das Finden optimaler Pläne garantieren; und es wurde ein prototypisches Assistenzsystem realisiert, das seinen Nutzer bei dem Aufbau einer komplexen Heimkinoanlage unterstützt.

## 1 Einleitung

Technische Systeme werden immer komplexer. Trotz der zunehmenden technischen Möglichkeiten sind Nutzer moderner technischer Geräte heutzutage immer noch oft auf sich allein gestellt. Meist muss der Nutzer erst erlernen, wie das entsprechende Gerät zu bedienen ist, wo er oder sie welche Funktion findet und was sie genau bewirkt. Unterstützung erhält man nur selten: Erklärungen sind meist nur in Form von unübersichtlichen Bedienungsanleitungen vorhanden, die vorgefertigte Texte darstellen und damit nicht auf die aktuelle Situation zugeschnitten sind. Sprachgesteuerte Assistenten verbessern bereits heute diese Situation, da sie ein einfaches Steuern der entsprechenden Geräte erlauben. *Companion-Technologie* jedoch hat das Ziel, technische Systeme jeder Art zu *kognitiven* technischen Systemen aufzuwerten, die ihre Nutzer durch kognitive Fähigkeiten wie Planen und Schlussfolgern anleiten und unterstützen [BW16, Bi16, BW17]. Ergänzt durch Technologien der Mensch-Maschine-Interaktion wird das generelle Bedienkonzept der Systeme nutzerfreundlicher gestaltet sowie Erklärbarkeit der eigenen Funktionalität und damit Transparenz erreicht. Die Dissertation legt die hierfür relevanten Grundlagen im Bereich der Handlungsplanung und zeigt, wie hierdurch flexible Assistenten realisiert werden können.

Durch den Einsatz von Handlungsplanung als Basistechnologie für Assistenzsysteme ergeben sich eine Vielzahl von Anforderungen an die Handlungsplanung. Grundvoraussetzung ist, dass der Planungsprozess möglichst schnell abgeschlossen ist, da nicht zu erwarten

[1] Englischer Originaltitel der Dissertation: „Hybrid Planning — From Theory to Practice“ [Be18]. Die zwölf Hauptbeiträge der kumulativen Dissertation sind insbesondere in der Conclusion als solche hervorgehoben.

[2] Institut für Künstliche Intelligenz, Ulm University, pascal.bercher@uni-ulm.de

ist, dass Nutzer ein System benutzen werden, welches lange Wartezeiten erfordert. Die dem Nutzer präsentierten Handlungsempfehlungen müssen außerdem plausibel sein. Dazu zählt, dass sie nicht „irgendwelche“ Lösungen darstellen, sondern möglichst *gute*, d.h. kostengünstige. Schließlich müssen die gefundenen Handlungsempfehlungen auch in einer intuitiven Reihenfolge dargestellt werden, so dass Handlungen, die zusammengehörige Teilprobleme adressieren, auch zusammen instruiert werden. Eine weitere Anforderung ist die Erklärbarkeit von Plänen, die dem Nutzer die Möglichkeit gibt, den Sinn einer dargestellten Handlungsempfehlung zu erfragen.

All diese Anforderungen werden durch die Dissertation adressiert. Zunächst wird der gewählte Planungsansatz, *hybrides Planen*, welches hierarchisches Planen mit nichtlinearem Planen verbindet, theoretisch analysiert. Hierfür wird ein neuer Formalismus vorgestellt, der sich bereits zu einem akzeptierten Standard etablieren konnte. Darauf basierend wurde eine Kategorisierung der untersuchten Problemklassen vorgenommen und neue Komplexitätsergebnisse in den resultierenden Subklassen erzielt, darunter auch erstmals Ergebnisse zum Planexistenz- und Planverifikationsproblem des hybriden Planens (siehe Kapitel 2). Es werden die ersten zulässigen Heuristiken für hybrides Planen und dessen Subklassen vorgestellt, die – erstmals – das Finden optimaler Lösungen für die jeweiligen Klassen durch heuristische Suche erlauben (siehe Kapitel 3). Als Proof-of-Concept wird ein prototypisches *Companion-System* vorgestellt, das seine Nutzer beim Aufbau einer Heimkinoanlage unterstützt und hierbei weitere Planungsfähigkeiten umsetzt wie Planerklärung, Planreparatur und die plausible Linearisierung nicht-linearer Pläne (siehe Kapitel 4). Kapitel 5 fasst die Hauptergebnisse nochmals zusammen.

## 2 Theoretische Grundlagen

Als Grundlage der Untersuchungen der vorgestellten Doktorarbeit wird ein hierarchischer Planungsansatz gewählt. Wie in Kapitel 4 näher ausgeführt, ergeben sich hieraus einige Vorteile im vorliegenden Kontext der Assistenzsysteme, die sich über alle Ebenen des Anwendungskontexts erstreckt: sie kann für Modellierungsunterstützung ausgenutzt werden, für eine zielgerichtetere Suche, für die Generierung von Erklärungen, und auch bei der Präsentation des Plans selbst.

Als Formalismus wird das sogenannte *hybride Planen* gewählt, ein hierarchischer Planungsansatz, der zwei traditionelle Ansätze kombiniert, nämlich das HTN-Planen mit dem POCL-Planen[1]. Das hybride Planen wurde bereits vor dieser Dissertation entwickelt [KMS98, BS01], aber die Dissertation stellt eine neue Formalisierung hiervon vor [Be16]. Diese Formalisierung geht in größten Teilen auf eine neue Formalisierung von HTN-Problemen zurück, welche ebenfalls ein Hauptbeitrag der Dissertation darstellt [GB11]. Diese neue Formalisierung dient seit ihrer Einführung 2011 bereits in zahlreichen wissenschaftlichen Publikationen als Grundlage (auch von verschiedenen externen Forschergruppen) und kann daher als neuer etablierter Standard betrachtet werden. Aus Platzgründen sei für eine Auflistung der entsprechenden Arbeiten (bis etwa Mitte 2017) auf Kapitel 2 der Dissertation verwiesen [Be18].

[1] Dabei steht *HTN* für *Hierarchical Task Network* und *POCL* für *Partial-Order Causal-Link*.

Da die Formalisierung des hybriden und HTN-Planens einen Hauptbeitrag der Dissertation darstellt, wird sie hier auszugsweise vorgestellt. Wir beginnen mit dem nicht-hierarchischen (d.h. *primitiven*) Anteil, dessen Formalisierung zu allgemein bekannten Grundlagen zählt. Ein nicht-hierarchisches Planungsproblem ist ein Tupel $\mathscr{P} = \langle \mathscr{D}, s_0, g \rangle$, bestehend aus der Planungsdomäne $\mathscr{D} = \langle V, A \rangle$, welche die zur Verfügung stehende Aktionsmenge beschreibt, dem Anfangszustand $s_0$ und der Zielzustandsbeschreibung $g$. Das Modell ist ein rein propositionales, das auf einer endlichen Menge von Zustandsvariablen $V$ beruht. Hierdurch wird implizit die Menge der möglichen Zustände $2^V$ definiert, wobei $s_0 \in 2^V$ dem Zustand vor der Ausführung eines Plans entspricht und $g = (g^+, g^-)$ mit $g^+, g^- \subseteq V$ die Menge der erwünschten Zielzustände beschreibt: Ein Zustand $s \in 2^V$ ist genau dann ein Zielzustand, wenn $s \supseteq g^+$ und $s \cap g^- = \emptyset$ gilt. Aktionen überführen Zustände ineinander. Sie sind Tupel $(pre^+, pre^-, eff^+, eff^-) \in (2^V)^4$, bestehend aus den positiven und negativen Vorbedingung (preconditions) $pre^+$ und $pre^-$ und den positiven und negativen Effekten $eff^+$ und $eff^-$. Eine solche Aktion ist genau dann in einem Zustand $s \in 2^V$ ausführbar, falls $s \supseteq pre^+$ und $s \cap pre^- = \emptyset$ gelten. Ist dies der Fall, ist der Folgezustand definiert als $s' := (s \setminus eff^-) \cup eff^+$. Eine Lösung eines nicht-hierarchischen Planungsproblems $\mathscr{P}$ ist dann eine Aktionssequenz, die in einen Zielzustand resultiert und deren Aktionen in ihrem jeweiligen Vorgängerzustand anwendbar sind, beginnend in $s_0$.

Unter hierarchischem Planen versteht man eine Problemklasse, die zusätzliche Anforderungen an Lösungen stellt. Es handelt sich also nicht um eine andere Art des Planens zum Finden von Lösungen, sondern um eine ausdrucksmächtigere Problemklasse, mit der man auch bestimmte Pläne *ausschließen* kann. Zu diesem Zweck werden die Aktionen der Planungsdomäne um eine Hierarchie ergänzt. In diesem neuen Kontext wird nun von Tasks gesprochen: *Primitive Tasks* sind die bereits bekannten Aktionen. *Abstrakte Tasks* hingegen müssen erst in primitive verfeinert werden. Hierfür enthält das Modell für jeden abstrakten Task eine Menge von *Dekompositionsmethoden*, welche Abbildungen auf vordefinierte *Tasknetzwerke* (kurz: *Tasknetze*) darstellen, die als Standardimplementierungen der abstrakten Tasks angesehen werden können [Be16]. Tasknetzwerke sind partiell geordnete Mengen von Tasks, die auch wiederum abstrakt sein können. In der vorgestellten Formalisierung des HTN-Planens sind sie wie folgt definiert:

**Definition 2.1.** *Ein* Tasknetzwerk $tn = (T, \prec, \alpha)$ *ist ein 3-Tupel, bestehend aus:*
- $T$*, einer endlichen Menge sogenannter Task-Identifizierungssysmbole,*
- $\prec \subseteq T \times T$*, einer strikten Partialordnung über* $T$ *und*
- $\alpha : T \to N$*, wobei* $N$ *eine endliche Menge von Tasknamen ist.*

Zu beachten ist, dass abstrakte Tasks nur aus einem Namen aus $N$ bestehen (denn ihr einziger Sinn ist ihre Verfeinerung durch eine Methode) und jeder primitive Task hat ebenfalls einen Namen, der als aussagekräftige Beschreibung statt des entsprechenden Aktionstupels verwendet werden kann. Eine Dekompositionsmethode ist nun ein 2-Tupel $(n, tn)$, wobei $n$ der Name eines abstrakten Tasks und $tn$ ein Tasknetzwerk, in das $n$ verfeinert werden kann. Gegeben ein Tasknetzwerk $tn = (T, \prec, \alpha)$, ein abstrakter Task $n = \alpha(t)$, mit $t \in T$ und eine Dekompositionsmethode $m = (n, tn')$, so ergibt sich aus der Anwendung von $m$ auf $tn$ ein neues Tasknetzwerk $tn''$, in welchem der Task $n$ durch $tn'$ ersetzt und die vorliegenden Ordnungsconstraints entsprechend vererbt werden [GB11].

Eine Lösung ist nun definiert als ein Tasknetzwerk, für das im Wesentlichen zwei Eigenschaften erfüllt sind: (1) es muss ausführbar sein, d.h. es besitzt eine Linearisierung seiner Aktionen, die den Startzustand in einen Zielzustand überführt und (2) es muss durch Anwendung einer Sequenz von Dekompositionsmethoden aus dem *initialen* Tasknetzwerk generierbar sein, welches Teil der hierarchischen Problembeschreibung $\mathscr{P}$ ist [GB11]. Dieser Formalismus besticht durch zwei wesentliche Punkte: er ist simplistisch, d.h. ohne komplizierte und unnötige Definitionen, und trotzdem vollständig und formal korrekt. Zweitens nennt er *explizit* die Lösungskriterien, was zuvor nur selten bis gar nie in ausreichendem Maße gemacht wurde, was zu einer Fehlinterpretation der Semantik und entsprechend zu einer unzureichenden oder gar falschen Einordnung in die Literatur führen konnte – insbesondere von hierarchischen Planungsalgorithmen [Be18].

Diese Formalisierung für HTN-Planen [GB11] wurde erweitert auf eine neue Formalisierung für hybrides Planen [Be16]. Im hybriden Planen werden Tasknetzwerke als Pläne bezeichnet und die Task-Identifizierungssymbole als Planschritte. Es unterscheidet sich vom HTN-Planen dadurch, dass auch abstrakte Tasks Vorbedingung und Effekte besitzen – dies erlaubt u.a. das direkte Schlussfolgern auf höheren Abstraktionsebenen. Weiter können sämtliche Pläne des Domänenmodells kausale Links enthalten. Ein kausaler Link ist ein Tupel $(ps, v, ps')$, der annotiert, dass die Vorbedingungsvariable $v$ des Planschritts $ps'$ durch den Planschritt $ps$ geschützt wird. Durch die angepassten Lösungskriterien wird garantiert, dass dieser „Schutz“ der kausalen Links auch eingehalten wird. Sowohl die Vorbedingungen und Effekte der abstrakten Tasks, als auch die in den Dekompositionsmethoden vorliegenden kausalen Links werden für die Definition sogenannter Legalitätskriterien benutzt, die bestimmen, ob Dekompositionsmethoden auch tatsächlich Implementierungen ihrer abstrakten Tasks sind [Be16]. In der Dissertation wird ein neues Implementierungskriterium vorgestellt, bestehende aus der Literatur in einem einheitlichen Formalismus dargestellt, ihr Einfluss auf Komplexitätseigenschaften dargelegt (siehe unten), und diskutiert, wie diese zur Modellierungsunterstützung genutzt werden können [Be16].

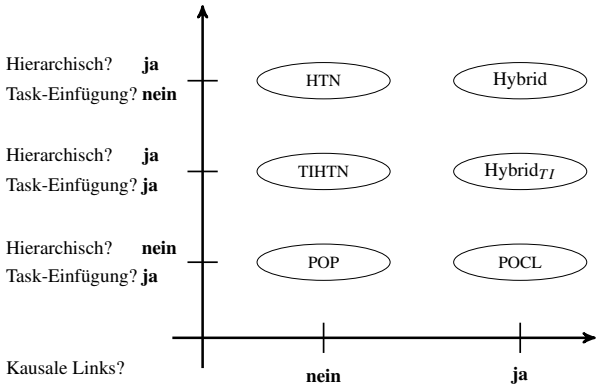



Abb. 1: Übersicht der untersuchten Problemklassen. Die Kategorisierung hängt davon ab, ob der initiale Plan oder irgend ein Plan irgend einer Methode kausale Links enthalten darf; davon, ob der initiale Plan abstrakte Tasks enthält; und davon, ob Task-Einfügung erlaubt ist.

Durch die vereinheitlichte Formalisierung wurde eine Kategorisierung der verschiedenen Problemklassen ermöglicht (für eine Übersicht siehe Abb. 1) und hierdurch eine systematische Betrachtung der theoretischen Eigenschaften der resultierenden Problemklassen.

Erst durch die neue Formalisierung des hybriden Planens war es möglich, eine Vielzahl der Komplexitätsergebnisse des HTN-Planens [ABA15a] auf hybrides Planen zu übertragen [Be16]. Zuvor waren *sämtliche* Komplexitätsergebnisse unbekannt. Es konnte bewiesen werden, dass das hybride Planen im allgemeinen Fall (d.h. ohne zusätzliche Einschränkungen wie beispielsweise total geordnete Pläne in den Methoden) exakt so ausdrucksmächtig ist wie HTN-Planen: Das Planexistenzproblem (*„Wie schwierig ist es, zu beweisen, ob eine Lösung existiert?“*) ist sowohl **unentscheidbar** als auch **semi-entscheidbar**. Weiter konnte durch eine allgemeine Reduktion gezeigt werden, dass viele bekannte Spezialfälle des hybriden Planens genauso schwierig sind wie im entsprechenden HTN-Fall. Membership-Beweise sind in Arbeit. Weiter wurde bewiesen, dass auch die Planverifikation (*„Ist der vorliegende Plan eine Lösung für das Planungsproblem?“*) genauso schwierig ist wie im HTN-Planen: Es ist **NP-vollständig**.

Auch im HTN-Planen wurden grundlegend neue und einflussreiche Ergebnisse erzielt. So erstmals systematisch untersucht, welchen Einfluss es nimmt, Tasknetze nicht nur durch Dekomposition abstrakter Tasks verfeinern zu dürfen – wie es im HTN-Planen üblich ist; vgl. Lösungskriterium (2) – sondern auch durch das Einfügen primitiver Aktionen, z.B., um Vorbedingungen anderer Aktionen zu schützen. Es wurde gezeigt, dass hierdurch die Komplexität des Planexistenzproblems von **unentscheidbar** auf **EXPSPACE** membership sinkt [GB11]. Hierdurch wurde eine prinzipiell neuartige Problemrelaxierung entdeckt, die Entscheidbarkeit herbeiführt. Die daraus resultierende Problemklasse, genannt TIHTN-Planen (HTN-Planen mit Task Insertion) wurde in der Folge für weitere theoretische Untersuchungen aufgegriffen [Al14, ABA15b, Hö16], auch von anderen Forschergruppen. Dabei hat sich diese neue Problemrelaxierung insbesondere als Grundlage für Heuristiken als praktikabel erwiesen [Al14, Be17a].

Auch im POCL-Planen wurden neue Grundlagenergebnisse erzielt. Für die Entwicklung von Heuristiken im POCL-Planen ist es essentiell, die Ursache für die Schwierigkeit des Planexistenzproblems für einen gegebenen partiellen Plan zu kennen. Dennoch waren vor Veröffentlichung unserer Ergebnisse nur unzureichende theoretische Untersuchungen dieses Problems vorliegend. Wir konnten zeigen, dass im allgemeinen Fall das Planexistenzproblem für einen POCL-Pan **PSPACE-vollständig** ist [Be18]. Sehr viel interessanter und daher bedeutender ist jedoch die Erkenntnis, dass das Problem noch immer **NP-vollständig** ist, wenn sämtliche Aktionen – sowohl im aktuellen Plan als auch im Domänenmodell – nur positive Vorbedingungen haben und die Aktionen im Domänenmodell darüber hinaus delete-relaxiert sind, d.h. auch keine negativen Effekte haben [Be13]. Dieses Ergebnis ist deswegen so interessant, da die analoge Frage im klassischen Planen, wo die Heuristik für eine Zustand statt für einen Plan evaluiert wird, ein in **P** entscheidbares Problem ist. Dies zeigt die besondere Rolle der partiellen Ordnung von Aktionen auf und gibt damit Hinweise für die Entwicklung von Heuristiken im POCL- und hybriden Planen.

## 3 Suche und Heuristiken

Der Einsatz von Planungstechnologie mit einem Menschen im Mittelpunkt lässt im Kontext des Lösens der vorliegenden Probleme zwei Anforderungen von zentraler Bedeu-

tung werden: Laufzeit und Qualität der Lösungen. Wenn ein Mensch von einem Assistenzsystem Unterstützung sucht, möchte er *umgehend* unterstützt werden – es ist kaum Verständnis für lange Wartezeiten zu erwarten. Gleichzeitig sollten die Lösungen auch plausibel sein. Ein Nutzer wäre sicherlich sehr verwundert darüber, eine Instruktion präsentiert zu bekommen, die durch die nächste oder übernächste Instruktion gleich wieder rückgängig gemacht wird. Es sollten daher *gute* Lösungen gefunden werden.

Die Dissertation stellt zunächst einen Algorithmus zum Lösen hybrider Planungsprobleme vor [BKB14]. Dieser Algorithmus ist eine Weiterentwicklung – mit beachtlichen Laufzeitverbesserungen – des bis dahin als State-of-the-Art geltenden Planungssystems zum dekompositionsbasierten Lösen von Planungsproblemen [Be18]. Da hybride Planungsprobleme eine Erweiterung von HTN-Problemen um POCL-Konzepte darstellen, kann der entsprechende Planer neben hybriden Problemen auch HTN-Probleme, TIHTN-Probleme, POCL-Probleme und klassische Planungsprobleme lösen. Zur Laufzeitverbesserung wurden eine Reihe neuer Heuristiken entwickelt.

Als Grundlage für Heuristiken im hierarchischen Planen wurde der Task-Dekompositions-Graph (TDG) entwickelt, eine Repräsentation der Aktionshierarchie [El12, Be17a]. Eine dieser Heuristiken nutzt *Landmarken*, um den Verfeinerungsaufwand eines Suchknotens zu schätzen [BKB14]. Landmarken sind Tasks, die durch Verfeinerung eines bestimmten abstrakten Tasks in jedem Fall in einen Plan eingefügt werden. Im nächsten Schritt wurden diese Heuristiken verallgemeinert, um auch in Abwesenheit von Landmarken informiert zu agieren: Sie nutzen die dem TDG zugrundeliegende UND/ODER-Struktur aus, indem der Aufwand von Tasks in derselben Methode aufsummiert wird, während der Aufwand eines abstrakten Tasks zur Minimierung über seine Methoden geschätzt werden kann [Be17a]. Diese Heuristiken zählen zu den ersten wohlinformierten Heuristiken im HTN- und hybriden Planen. Zuvor wurde für die Heuristikberechnung entweder die Aktionshierarchie nicht hinreichend betrachtet oder es wurden gleich gänzlich keine Heuristiken zur Suchsteuerung verwendet; stattdessen musste eine Suchstrategie in das Modell kodiert werden [Be18]. Die vorgestellten Heuristiken sind außerdem zulässig, d.h. mit einem entsprechenden Algorithmus wie $A^*$ garantieren sie das Finden optimaler Lösungen. Dies war vor dieser Arbeit nur durch hartkodierte Heuristikinformation oder durch uninformierte Suche möglich (d.h. mit uniformer Kostensuche).

Auch im POCL-Planen wurden neue Heuristiken entwickelt. Die erste Heuristik, *sampleFF*, approximiert das **NP-schwierige** Entscheidungsproblem der delete-relaxierten Probleme (siehe letztes Kapitel) durch das Samplen einer fixen Anzahl von Linearisierungen eines gegebenen Plans [Be13]. Aufgrund hoher Laufzeiten ist diese Heuristik nicht kompetitiv mit dem State-of-the-Art in Bezug auf gelöste Probleminstanzen auf einem Standard-Benchmarkset. Da sie aber mehr Informationen über die vorliegenden Pläne nutzt als andere Heuristiken, war sie die einzige, die ein unlösbares Problem als unlösbar beweisen konnte. Eine zweite „Heuristik“ beruht auf einer Kodierung eines Plans in ein neues Modell [BGB13]. Hierdurch kann sie Heuristiken aus dem klassischen Planen zugänglich machen, die den Zielabstand für *Zustände* schätzt. Zulässige zustandsbasierte Heuristiken sind aufgrund der Kodierungseigenschaften auch im POCL-Planen zulässig und erlauben damit – zum ersten Mal im POCL-Planen – das heuristische Suchen *optimaler* Pläne.

## 4 Praktische Anwendung

Das schnelle Finden guter oder optimaler Lösungen ist für die Nutzung eines tatsächlich eingesetzten Assistenzsystems zwar eine wichtige Voraussetzung, doch in der Praxis ergeben sich weitere Anforderungen an die Handlungsplanung. Wir konnten insbesondere die Fähigkeiten der Planerklärung, Planausführungsüberwachung mit Planreparatur sowie die adäquate Präsentation der Pläne selbst identifizieren [Bi11, Be17b].

Während die Planreparatur und die Fähigkeiten zur Planerklärung bereits vor dieser Dissertation durch Kollegen erforscht wurden, ist die sogenannte *Planlinearisierung* ein weiterer Beitrag der Dissertation [Be17b]. Hierbei handelt es sich um domänenunabhängige Strategien, die nutzerfreundliche Linearisierungen erzeugen. Als Eingabe dient ein Lösungsplan, also ein nicht-linearer Plan, der garantiert, dass jede Linearisierung ausführbar ist und sämtliche Nutzerziele erfüllt. Dennoch könnten einige Linearisierungen für Menschen (un)intuitiver sein als andere. Es ist die Aufgabe der Linearisierungskomponente, solche Linearisierungen zu bevorzugen, die menschliche Nutzer intuitiv erscheinen.

Ein weiterer Beitrag ist die Operationalisierung der genannten nutzerzentrierten Planungsfähigkeiten im Kontext von *Companion-Systemen*. Hierzu wurde mit zahlreichen Kollegen verschiedener Institute der Universität Ulm ein prototypisches *Companion-System* realisiert. Das System unterstützt seinen Nutzer bei der Verkabelung einer Heimkinoanlage und realisiert hierfür alle genannten nutzerzentrierten Planungsfähigkeiten [Be14, Be15, Be17c]. Eine portable Version des Assistenten [Be15] wurde an verschiedenen KI-Konferenzen vorgestellt, darunter an der ICAPS ’14 und der AAAI ’15. An der AAAI wurde das entsprechende Papier als eines von fünf Papieren ausgewählt, das an einer Pressekonferenz vorgestellt werden durfte, um es so einer breiten Öffentlichkeit zugänglich zu machen. Nominierungs- und Auswahlkriterium war der Nutzen für die Gesellschaft.

Im gewählten Beispielszenario des Assistenten besteht diese Heimkinoanlage aus einem Fernseher, einem Audio/Video-Verstärker, an welchen die Boxen angeschlossen sind, sowie einem Satellitenempfänger und einem Blu-ray-Player. Zur Verkabelung stehen dem Nutzer eine Vielzahl von Kabeln und Adaptern zur Verfügung. Diese Aufgabe wurde als Planungsproblem modelliert. Neben diesem formalen Modell der verwendeten Hardware (Geräte, Kabel, etc.) liegen dem System zahlreiche Bilder und Videos hiervon vor. Diese werden als Grundlage verwendet, um dem Nutzer eine detaillierte Schritt-für-Schritt-Anleitung zu präsentieren, die zuvor vollautomatisch basierend auf den Modellen erstellt wurde. Während das System zwar für ein bestimmtes Anwendungsszenario implementiert wurde, ist die zugrundeliegende Architektur domänenunabhängig und in vielen Szenarien einsetzbar [Be14].

## 5 Conclusion

Die Dissertation beschäftigt sich mit dem hybriden Planungsformalismus – von der Theorie zur Praxis. Hybrides Planen ist eine Fusionierung von HTN-Planen (ein hierarchischer Planungsansatz) mit POCL-Planen (ein nicht-lineares Planungsverfahren, das auf der ex-

pliziten Repräsentation von Kausalzusammenhängen in Plänen basiert). Die Arbeit untersucht, ob und wie der Formalismus als Grundlage für flexible und intelligente Assistenten eingesetzt werden kann. Im Kontext solcher *Companion-Systeme* [Bi16] werden von Theorie bis Praxis verschiedene Ergebnisse erzielt, die nachfolgend dargestellt werden.

Als eines der Hauptergebnisse wurde eine neue Formalisierung von HTN-Planungsproblemen eingeführt [GB11]. Sie ist besonders gut zum Führen von Beweisen formaler Eigenschaften geeignet und wurde insbesondere zu diesem Zweck bereits oft aufgegriffen [Be18]. Durch ihre Erweiterung auf hybrides Planen konnten außerdem die ersten Komplexitätsergebnisse für hybrides Planen nachgewiesen werden [Be16]. Es ist im allgemeinen Fall **unentscheidbar** und auch in vielen Spezialfällen (z.B. wenn alle Pläne total geordnet sind) mindestens so schwierig wie HTN-Planen. Die Verifikation, ob ein Plan eine Lösung ist, ist **NP-vollständig**, wie auch im HTN-Planen. Neue Komplexitätsergebnisse wurden außerdem in den Bestandteilen des hybriden Formalismus erzielt. Es konnte gezeigt werden, dass die Fähigkeit der Task-Einfügung (hier dürfen beliebig Tasks in Tasknetze eingefügt werden; nicht nur durch die Dekomposition von abstrakten Tasks wie sonst üblich) das **unentscheidbare** HTN-Planexistenzproblem **entscheidbar** macht [GB11]. Hierdurch wurde eine prinzipiell neuartige Form der Problemrelaxierung geschaffen, die einerseits theoretisch von Bedeutung ist, andererseits aber auch praktische Konsequenzen hat, z.B. zur Heuristikberechnung. Für das POCL-Planen wurden die ersten Komplexitätsergebnisse nachgewiesen [Be13]. Es wurde insbesondere bewiesen, dass Delete-Relaxierung der Aktionen im Domänenmodell das sonst **PSPACE-vollständige** Planexistenzproblem **NP-vollständig** macht.

Zur Lösung aller abgedeckten Problemklassen wurde ein effizienter heuristischer Suchalgorithmus vorstellt [BKB14]. Zum Lösen von POCL-Problem wurden neue Heuristiken entwickelt [Be13, BGB13], darunter ein Ansatz, der basierend auf einer Problemtransformation Heuristiken aus dem zustandsbasierten Planen zugänglich macht [BGB13]. Durch ihn wurden die ersten zulässigen Heuristiken im POCL-Planen erzielt, welche das Finden optimaler Lösungen ermöglicht. Es wurden auch neue Suchstrategien und Heuristiken für das HTN- und hybride Planen entwickelt [El12, Be17a]. Eine dieser Heuristiken ist die erste zulässige Heuristik für diese Klassen und erlaubt daher erstmals das Finden optimaler Lösungen durch heuristische Suche [Be17a].

Es wurde ein prototypisches *Companion-System* entwickelt [Be14, Be15, Be17c][2], das als Proof-of-Concept veranschaulicht, wie nutzerzentrierte Planungsfähigkeiten wie Reparatur, Erklärung und Planlinearisierung in einem praktisch eingesetzten System operationalisiert werden können und hierdurch flexibel seine Nutzer unterstützt [Bi11, Be17b].

**Danksagung**

Ich danke meiner Doktormutter Prof. Dr. Susanne Biundo-Stephan, mir die Möglichkeit zur Promotion in einem gesellschaftlich relevanten Gebiet zu ermöglicht zu haben, wel-

[2] Das zitierte Buchkapitel [Be17c] ist in der Dissertation nicht als Hauptbeitrag hervorgehoben. Es wird hier insbesondere zitiert, da es die Einzelbeiträge der verschiedenen Autorengruppen am Gesamtsystem darstellt.

ches trotz des praktischen Anwendungskontexts auch intensive theoretische Untersuchungen erlaubt. Der Universität Ulm danke ich für die Nominierung zum GI-Dissertationspreis.

## Literatur

[ABA15a] Alford, Ron; Bercher, Pascal; Aha, David: Tight Bounds for HTN Planning. In: Proc. of the 25th Int. Conf. on Automated Planning and Scheduling (ICAPS). AAAI Press, S. 7–15, 2015.

[ABA15b] Alford, Ron; Bercher, Pascal; Aha, David: Tight Bounds for HTN planning with Task Insertion. In: Proc. of the 25th Int. Joint Conf. on AI (IJCAI). AAAI Press, S. 1502–1508, 2015.

[Al14] Alford, Ron; Shivashankar, Vikas; Kuter, Ugur; Nau, Dana: On the Feasibility of Planning Graph Style Heuristics for HTN Planning. In: Proc. of the 24th Int. Conf. on Automated Planning and Scheduling (ICAPS). AAAI Press, S. 2–10, 2014.

[Be13] Bercher, Pascal; Geier, Thomas; Richter, Felix; Biundo, Susanne: On Delete Relaxation in Partial-Order Causal-Link Planning. In: Proc. of the 2013 IEEE 25th Int. Conf. on Tools with AI (ICTAI). IEEE Computer Society, S. 674–681, 2013.

[Be14] Bercher, Pascal; Biundo, Susanne; Geier, Thomas; Hörnle, Thilo; Nothdurft, Florian; Richter, Felix; Schattenberg, Bernd: Plan, Repair, Execute, Explain - How Planning Helps to Assemble your Home Theater. In: Proc. of the 24th Int. Conf. on Automated Planning and Scheduling (ICAPS). AAAI Press, S. 386–394, 2014.

[Be15] Bercher, Pascal; Richter, Felix; Hörnle, Thilo; Geier, Thomas; Höller, Daniel; Behnke, Gregor; Nothdurft, Florian; Honold, Frank; Minker, Wolfgang; Weber, Michael; Biundo, Susanne: A Planning-based Assistance System for Setting Up a Home Theater. In: Proc. of the 29th AAAI Conference on AI (AAAI). AAAI Press, S. 4264–4265, 2015.

[Be16] Bercher, Pascal; Höller, Daniel; Behnke, Gregor; Biundo, Susanne: More than a Name? On Implications of Preconditions and Effects of Compound HTN Planning Tasks. In: Proc. of the 22nd Europ. Conf. on AI (ECAI). IOS Press, S. 225–233, 2016.

[Be17a] Bercher, Pascal; Behnke, Gregor; Höller, Daniel; Biundo, Susanne: An Admissible HTN Planning Heuristic. In: Proc. of the 26th Int. Joint Conf. on AI (IJCAI). AAAI Press, S. 480–488, 2017.

[Be17b] Bercher, Pascal; Höller, Daniel; Behnke, Gregor; Biundo, Susanne: User-Centered Planning. In Biundo; Wendemuth, A. [BW17], 2017. Kapitel 5, S. 79–100.

[Be17c] Bercher, Pascal; Richter, Felix; Hörnle, Thilo; Geier, Thomas; Höller, Daniel; Behnke, Gregor; Nothdurft, Florian; Honold, Frank; Schüssel, Felix; Reuter, Stephan; Minker, Wolfgang; Weber, Michael; Dietmayer, Klaus; Biundo, Susanne: Advanced User Assistance for Setting Up a Home Theater. S. 485–491. In Biundo; Wendemuth, A. [BW17], S. 485–491, 2017. Kapitel 24, S. 485–49.

[Be18] Bercher, Pascal: Hybrid Planning — From Theory to Practice. Dissertation, Ulm University, 2018. doi: 10.18725/OPARU-5242.

[BGB13] Bercher, Pascal; Geier, Thomas; Biundo, Susanne: Using State-Based Planning Heuristics for Partial-Order Causal-Link Planning. In: Advances in AI, Proceedings of the 36th German Conference on AI (KI). Springer, S. 1–12, 2013.

[Bi11] Biundo, Susanne; Bercher, Pascal; Geier, Thomas; Müller, Felix; Schattenberg, Bernd: Advanced user assistance based on AI planning. Cognitive Systems Research, 12(3-4):219–236, April 2011. Special Issue on Complex Cognition.

[Bi16] Biundo, Susanne; Höller, Daniel; Schattenberg, Bernd; Bercher, Pascal: Companion-Technology: An Overview. Künstliche Intelligenz, 30(1):11–20, 2016.

[BKB14] Bercher, Pascal; Keen, Shawn; Biundo, Susanne: Hybrid Planning Heuristics Based on Task Decomposition Graphs. In: Proc. of the 7th Annual Symposium on Combinatorial Search (SoCS). AAAI Press, S. 35–43, 2014.

[BS01] Biundo, Susanne; Schattenberg, Bernd: From Abstract Crisis to Concrete Relief – A Preliminary Report on Combining State Abstraction and HTN Planning. In: Proc. of the 6th Europ. Conf. on Planning (ECP). AAAI Press, S. 157–168, 2001.

[BW16] Biundo, Susanne; Wendemuth, Andreas: *Companion*-Technology for Cognitive Technical Systems. Künstliche Intelligenz, 30(1):71–75, 2016.

[BW17] Biundo, Susanne; Wendemuth, Andreas: Companion Technology – A Paradigm Shift in Human-Technology Interaction. Cognitive Technologies. Springer, 2017.

[El12] Elkawkagy, Mohamed; Bercher, Pascal; Schattenberg, Bernd; Biundo, Susanne: Improving Hierarchical Planning Performance by the Use of Landmarks. In: Proc. of the 26th AAAI Conf. on Artificial Intelligence (AAAI). AAAI Press, S. 1763–1769, 2012.

[GB11] Geier, Thomas; Bercher, Pascal: On the Decidability of HTN Planning with Task Insertion. In: Proc. of the 22nd Int. Joint Conf. on AI (IJCAI). AAAI Press, S. 1955–1961, 2011.

[Hö16] Höller, Daniel; Behnke, Gregor; Bercher, Pascal; Biundo, Susanne: Assessing the Expressivity of Planning Formalisms through the Comparison to Formal Languages. In: Proc. of the 26th Int. Conf. on Automated Planning and Scheduling (ICAPS). AAAI Press, S. 158–165, 2016.

[KMS98] Kambhampati, Subbarao; Mali, Amol; Srivastava, Biplav: Hybrid Planning for Partially Hierarchical Domains. In: Proc. of the 15th Nat. Conf. on AI (AAAI). AAAI Press, S. 882–888, 1998.

**Pascal Bercher**, Jahrgang '82, studierte von 2002 bis 2009 an der Albert-Ludwigs-Universität Freiburg im Breisgau Informatik (Diplom) mit Spezialisierung Künstliche Intelligenz und Nebenfach Kognitionswissenschaft. Mitte 2009 wechselte er zur Promotion bei Prof. Dr. Susanne Biundo-Stephan an das Institut für Künstliche Intelligenz der Universität Ulm. Hier arbeitete er im Sonderforschungsbereich SFB/TRR 62 *„Eine Companion-Technologie für kognitive technische Systeme“*. Er untersuchte, ob und wie man durch Planungstechnologie Assistenzsysteme realisieren kann. Schwerpunkte setzte er bei der Erforschung theoretischer Grundlagen sowie von Heuristiken zur Laufzeitverbesserung der eingesetzten Suchalgorithmen. Seit seiner Promotion Ende 2017 beschäftigt sich Herr Bercher im Rahmen seiner Habilitation am selben Institut weiter mit der Verbindung von Theorie und Praxis im Kontext von intelligenten Assistenzsystemen. Seit Ende 2016 koordiniert er in diesem Kontext die Ulmer Arbeiten des aus dem SFB entstandenen Transferprojekts *„Do it yourself, but not alone: Companion-Technologie für die Heimwerkerunterstützung“*, welches in Kooperation mit dem Projektpartner Robert Bosch GmbH durchgeführt wird.

# Über das Lösen nichtgrundierter Answer-Set-Programme unter Berücksichtigung der Baumweite[1]

Bernhard Bliem[2]

**Abstract:** Viele wichtige Probleme, die in der Praxis auftreten, sind NP-schwer. Ein etablierter Ansatz, um eine Vielzahl solcher Probleme zu lösen, ist die deklarative Sprache Answer Set Programming (ASP). Allerdings ist dies für einige NP-schwere Probleme trotz der beeindruckenden Effizienz von ASP-Solvern nicht praktikabel. Ein Hoffnungsschimmer ist die Beobachtung, dass in der Praxis auftretende Probleminstanzen oft kleine Baumweite aufweisen. Es konnte nämlich beobachtet werden, dass moderne ASP-Solver effizienter sind, wenn ihre Eingabe kleine Baumweite hat. Leider ist die Eingabe dieser Solver üblicherweise nicht die Probleminstanz selbst, sondern eine durch sogenanntes Grundieren gewonnene Zwischeninstanz, und dieser Prozess kann die Baumweite drastisch vergrößern. Der Einfluss des Grundierens auf die Baumweite wurde bisher nicht hinreichend verstanden.

In der Dissertation klären wir die Frage, unter welchen Umständen grundiert werden kann ohne die Baumweite deutlich zu erhöhen. Dazu definieren wir Klassen von ASP-Programmen und beweisen, dass das Grundieren solcher Programme zusammen mit der Eingabe nicht mit einem übermäßigen Anstieg der Baumweite einhergeht. Können wir ein Problem in einer solchen Klasse ausdrücken, profitieren wir somit automatisch von der augenscheinlichen „Empfindlichkeit“ von ASP-Solvern gegenüber Baumweite. Außerdem präsentieren wir Fortschritte in einer algorithmischen Methodik für das explizite Ausnutzen beschränkte Baumweite. Hier zielen wir insbesondere auf Probleme ab, welche Teilmengenminimierung erfordern, wie es für zahlreiche Probleme auf der zweiten Stufe der Polynomiellen Hierarchie der Fall ist. Schließlich klären wir einige seit langem unbeantwortete Fragen über die Komplexität von Allianzproblemen in Graphen.

## 1 Einführung

Answer Set Programming (ASP) hat sich zu einem äußerst beliebten Paradigma zum Lösen schwieriger Berechnungsprobleme entwickelt. Es bietet eine leicht verwendbare Sprache, welche prägnante Problemspezifikationen ermöglicht, und kann mit hocheffizienten Systemen aufwarten. Um mit ASP ein Problem zu lösen, spezifiziert man dieses, üblicherweise unabhängig von den Probleminstanzen, als eine Menge von *Regeln*, welche von den Lösungen zu erfüllende Bedingungen formalisieren. Um eine solche Menge von Regeln, genannt *ASP-Programm*, zu lösen, rufen ASP-Systeme üblicherweise zuerst einen *Grounder* auf, der ein äquivalentes *grundiertes*, d. h. variablenfreies, Programm ausgibt. Anschließend wird dieses grundierte Programm an einen *Solver* weitergereicht, der die *Answer Sets*, d. h. die Lösungen, des Programms berechnet. Wir verzichten auf eine Darstellung der Syntax und Semantik und verweisen hierzu auf die Einführung [BET11]. Wir illustrieren die Verwendung von ASP lediglich anhand des folgenden Beispiels.

**Beispiel 1.** Das Dreifärbbarkeitsproblem lässt sich in ASP wie folgt kodieren:

[1] Originaltitel: *Treewidth in Non-Ground Answer Set Solving and Alliance Problems in Graphs*

[2] TU Wien, bliem@dbai.tuwien.ac.at

```
rot(X) | gruen(X) | blau(X) :- knoten(X).
:- kante(X,Y), rot(X), rot(Y).
:- kante(X,Y), gruen(X), gruen(Y).
:- kante(X,Y), blau(X), blau(Y).
```

Großbuchstaben, wie $X$ und $Y$ in diesem Beispiel, sind Variablen und können durch Konstanten, die in der Eingabe vorkommen, instantiiert werden. In der ersten Zeile raten wir eine Farbe für jeden Knoten. Die *Constraints* in den verbleibenden Zeilen stellen sicher, dass benachbarte Knoten unterschiedliche Farben haben. Wir können dieses Programm zusammen mit einem Eingabegraphen, welcher mithilfe der Prädikate `knoten` und `kante` spezifiziert wird, einem ASP-System geben, welches dann alle gültigen Dreifärbungen des Graphen berechnet.

Wenn wir den gerichteten Graphen, der aus lediglich zwei Knoten $a, b$ und einer Kante $(a, b)$ besteht, auf diese Weise spezifizieren und zusammen mit dem obigen Programm einem Grounder vorlegen, könnte dieser folgendes äquivalente grundierte Programm produzieren:

```
knoten(a). knoten(b). kante(a,b).
rot(a) | gruen(a) | blau(a) :- knoten(a).
rot(b) | gruen(b) | blau(b) :- knoten(b).
:- kante(a,b), rot(a), rot(b).
:- kante(a,b), gruen(a), gruen(b).
:- kante(a,b), blau(a), blau(b).
```

Wie in diesem Beispiel führen Grounder in der Praxis diverse Optimierungen durch anstatt auf naive Weise alle Variablen durch alle möglichen Konstanten zu ersetzen. So haben wir die Regel `:- kante(b,b), blau(b), blau(b)` nicht angeführt, weil der Eingabegraph keine Kante $(b, b)$ enthält und `kante` ein *extensionales Prädikat* ist. Das bedeutet, dass dieses Prädikat vom Programm nicht hergeleitet wird sondern nur in der Eingabe vorkommt.

Geben wir dieses grundierte Programm an einen ASP-Solver, so gibt dieser als Answer Sets genau die Dreifärbungen des Graphen aus. (Da die Semantik von ASP eine gewisse Minimalität vorsieht, wird vermieden, dass ein Knoten mehrere Farben bekommt.) △

Trotz der mittlerweile beachtlichen Effizienzfortschritte von ASP-Systemen haben diese weiterhin Schwierigkeiten mit einigen herausfordernden Problemen. Dies ist nicht immer nur eine Frage der Komplexität im Sinne der klassischen Komplexitätstheorie. Interessanterweise kommt es mitunter vor, dass ASP-Systeme auf einem Problem sehr performant sind, während sie für ein anderes Problem von gleicher Komplexität deutlich länger brauchen. Oft ist die klassische Komplexitätstheorie daher nur von beschränkter Nützlichkeit, um die Leistung von ASP-Systemen in der Praxis zu erklären. In solchen Fällen kann es erkenntnisbringend sein, die *parametrisierte* Komplexität der Probleme zu betrachten. Mithilfe dieser Theorie lässt sich die Komplexität eines Problems nicht nur hinsichtlich der Eingabegröße untersuchen, sondern auch in Bezug auf andere Parameter.

In dieser Arbeit interessieren wir uns besonders für die Auswirkung des strukturellen Parameters *Baumweite* auf die Leistung von ASP-Solvern. Die Grundidee: Je kleiner die Baumweite eines Graphen ist, desto mehr ähnelt dieser einem Baum. Es ist altbekannt,

dass viele NP-schwere Graphenprobleme effizient lösbar werden, wenn wir die Eingabe auf Bäume beschränken, und es hat sich herausgestellt, dass dies bei vielen wichtigen Problemen sogar für die allgemeinere Klasse von Graphen beschränkter Baumweite gilt [ALS91].

Tatsächlich sind viele NP-schwere Probleme *FPT* („fixed-parameter tractable“) bzgl. des Parameters Baumweite, d. h. sie können in der Zeit $\mathcal{O}(f(k) \cdot n^c)$ gelöst werden, wobei $f$ eine beliebige berechenbare, nur von der Baumweite $k$ abhängige, Funktion ist, $n$ die Eingabegröße bezeichnet und $c$ eine beliebige Konstante ist. Solche Algorithmen sind üblicherweise für kleine Werte der Baumweite $k$ sehr effizient. Glücklicherweise konnte beobachtet werden, dass in der Praxis auftretende Instanzen üblicherweise kleine Baumweite aufweisen [Bo93]. Baumweite ist nicht nur für Graphenprobleme relevant sondern ebenso auf Instanzen verschiedenartigster Probleme anwendbar, indem man eine passende Repräsentation der Instanz als Graph wählt.

Es gab bereits einige Untersuchungen zu Baumweite in Bezug auf grundiertes ASP. Ein wichtiges Ergebnis ist der Algorithmus von [JPW09], der für grundierte ASP-Programme von beschränkter Baumweite in linearer Zeit entscheiden kann, ob das Programm eine Lösung hat. Dieser Algorithmus verwendet eine Technik namens *Dynamische Programmierung auf Baumzerlegungen*, welche sehr gängig für Algorithmen ist, die kleine Baumweite ausnutzen. Der Algorithmus von [JPW09] wurde auch implementiert und als Solver für grundiertes ASP vorgestellt [Mo10]. Für einige Probleme war dieser auf Dynamischer Programmierung basierende Solver in der Lage, die Leistung moderner ASP-Solver zu übertreffen, solange die Instanzen sehr groß waren und eine sehr kleine Baumweite hatten.

## 2 Problemstellung

Obwohl die ermutigenden Ergebnisse von [Mo10] bestätigten, dass kleine Baumweite erfolgreich für ASP-Solving unter „Laborbedingungen“ verwendet werden kann, waren die nötigen Einschränkungen hinsichtlich Problemen und Instanzen, um diesen Ansatz effizient anwenden zu können, zu schwerwiegend für die meisten praktischen Anwendungen. Die größten Hindernisse, diesen Ansatz für eine breite Vielfalt an Problemen nutzbar zu machen, waren die Tatsachen, dass einerseits naive Dynamische Programmierung unter einem enormen Overhead leidet (besonders bezüglich Speicherbedarf), und dass andererseits moderne ASP-Solver dermaßen effizient sind, dass sich die theoretische Überlegenheit des Dynamische-Programmierung-Algorithmus nur bei Instanzen gewaltiger Größe bezahlt macht.

Experimente in [Bl17] wiesen darauf hin, dass moderne ASP-Solver „empfindlich“ für die Baumweite ihrer Eingabe sind insofern als kleinere Baumweite stark mit höherer Performanz korreliert. Diese Beobachtungen lassen interessante Forschungsaufgaben erahnen. Insbesondere zwei Ansätze erscheinen vielversprechend, um kleine Baumweite erfolgreich für ASP-Solving in der Praxis nutzbar zu machen:

1. Die erste Forschungsaufgabe besteht darin, die auf Dynamischer Programmierung basierende Methodik zu verbessern, um deren Overhead und redundante Berechnungen zu vermindern.
   Verglichen mit anderen Problemen sind diese Punkte für das Lösen von grundiertem ASP besonders schwerwiegend, weil die entsprechenden Berechnungsprobleme (unter gängigen komplexitätstheoretischen Annahmen) noch schwieriger sind als NP. (Zu entscheiden, ob ein grundiertes ASP-Programm mit Disjunktionen ein Answer Set besitzt, befindet sich nämlich auf der zweiten Stufe der Polynomiellen Hierarchie.) Diese hohe Komplexität von grundiertem ASP widerspiegelt sich im Dynamische-Programmierung-Algorithmus [JPW09], welcher zuerst eine Brute-Force-Methode anwendet, um alle Modelle aller Teile des zerlegten Programms zu finden, und anschließend Brute Force erneut verwendet, um *für jedes solche partielle Modell* alle möglichen Gegenbeispiele zu finden, die zum Verwerfen des Kandidaten führen.
   Dieses Muster tritt zudem häufig in Dynamische-Programmierung-Algorithmen für andere Probleme auf, wo nach Lösungen gesucht wird, welche eine gewisse Teilmengenminimalität erfüllen (d. h. keine echte Teilmenge darf die Lösungsbedingungen erfüllen). Neben grundiertem ASP ist dies beispielsweise der Fall für das Problem, teilmengenminimale Modelle einer aussagenlogischen Formel zu finden. Im Allgemeinen treten Probleme mit Teilmengenminimierung recht häufig in z. B. der KI auf. Algorithmen für solche Probleme speichern üblicherweise eine große Anzahl redundanter Objekte, da die Teilmengen, die einen Lösungskandidaten entkräften, ihrerseits Lösungskandidaten sind. Weiters enthalten die Spezifikationen solcher Algorithmen selbst Redundanzen, da die potentiellen Gegenbeispiele normalerweise auf nahezu gleiche Weise behandelt werden wie die Lösungskandidaten.

2. Die zweite Forschungsaufgabe besteht darin, ASP zu lösen und dabei nicht Dynamische Programmierung durchzuführen sondern stattdessen kleine Baumweite *implizit* auszunutzen, indem man sich auf die (von Experimenten in [Bl17] gestützte) Annahme verlässt, dass moderne ASP-Solver effizienter sind, wenn man ihnen grundierte Programme von kleiner Baumweite vorlegt.
   Da Probleme üblicherweise in nichtgrundiertem ASP kodiert werden, ist hier das Forschungsziel, zu untersuchen, welche Kodierungstechniken in nichtgrundiertem ASP die Baumweite des grundierten Programms erheblich vergrößern, verglichen mit der Baumweite der Eingabe.

Neben der Nutzung von Baumweite für das Lösen von ASP sind wir darüber hinaus an einigen Varianten eines Graphenproblems namens SECURE SET [BDH07] interessiert. Es gehört zur Klasse der sogenannten *Allianzprobleme*, welche nach einer Gruppe von Knoten fragen, die einander auf eine bestimmte Weise aushelfen können. Zu praktischen Anwendungen von Allianzproblemen zählen das Finden von Gruppen von Webseiten, die Gemeinschaften bilden [Fl02] oder das Aufteilen von Ressourcen in einem Computernetzwerk sodass gleichzeitige Anfragen erfüllt werden können [HHH03]. Wir nennen eine Menge $S$ von Knoten eines Graphen *gesichert*, wenn jede Teilmenge von $S$ mindestens so viele Nachbarn in $S$ hat wie Nachbarn außerhalb von $S$. (Formal: Die Ungleichung $|N[X] \cap S| \geqslant |N[X] \setminus S|$ muss für jede Teilmenge $X$ von $S$ gelten, wobei $N[X]$ die *geschlossene Nachbarschaft* von $X$ ist, d. h. die Knoten in $X$ und deren Nachbarn.) Das SECURE

SET Problem fragt, ob ein gegebener Graph eine gesicherte Knotenmenge von höchstens einer gegebenen Größe enthält.

Der Grund, warum wir uns mit SECURE SET beschäftigen, ist, dass dieses Problem besonders für die ASP-Forschung sehr interessante Eigenschaften aufweist: Versuche, dieses Problem in ASP auszudrücken, führten zu äußerst komplizierten Spezifikationen, welche darauf hinweisen, dass SECURE SET womöglich die volle Ausdrucksstärke von ASP benötigt [Ab15]. Es ist jedoch leider unklar, ob dies tatsächlich der Fall ist, da die Komplexität des Problems ungeklärt geblieben ist, obwohl dieses bereits 2007 vorgestellt wurde [BDH07].

Eine der Varianten von SECURE SET, welche wir in der vorliegenden Arbeit behandeln, ist das DEFENSIVE ALLIANCE Problem. Hier suchen wir nach Knotenmengen $S$, wo für jedes *Element* $v \in S$ die Ungleichung $|N[v] \cap S| \geqslant |N[v] \setminus S|$ erfüllt ist. Dieses Problem hat in der Fachliteratur beachtliche Aufmerksamkeit genossen [FRV14]. Es ist als NP-vollständig bekannt, doch seine Komplexität parametrisiert durch Baumweite ist bislang offen geblieben.

## 3 Forschungsergebnisse

Unsere Ergebnisse können in drei Gruppen eingeteilt werden: Erstens präsentieren wir Fortschritte bei der Dynamischen Programmierung; zweitens definieren wir Klassen von nichtgrundiertem ASP, von denen wir zeigen, dass das Grundieren hier beschränkte Baumweite der Eingabe erhalten kann; drittens stellen wir Komplexitätsresultate und Algorithmen für Allianzprobleme in Graphen vor.

### 3.1 Fortschritte bei der Dynamischen Programmierung

Wir stellen eine fortgeschrittene Variante der Dynamischen Programmierung vor, die auf Teilmengenminimierung beinhaltende Probleme abzielt. Genauer gesagt formalisieren wir, wie für jedes Problem $P$, dessen Lösungen exakt die teilmengenminimalen Lösungen eines Grundproblems $G$ sind, ein Dynamische-Programmierung-Algorithmus für $G$ automatisch in einen Dynamische-Programmierung-Algorithmus für $P$ umgewandelt werden kann. Wir beweisen, dass die Laufzeit des resultierenden Algorithmus linear auf Instanzen beschränkter Baumweite ist, sofern dies für den Grundalgorithmus der Fall ist. Weiters zeigen wir, dass der resultierende Algorithmus korrekt ist, wenn der Grundalgorithmus korrekt ist und, sinngemäß, ausschließlich Teillösungen berechnet, die keine Entscheidungen, welche weiter unten in der Baumzerlegung getroffen worden sind, „rückgängig macht“. Der resultierende Algorithmus hat zwei Vorteile verglichen mit einem naiven Dynamische-Programmierung-Algorithmus, der $P$ direkt löst: Erstens ist er üblicherweise einfacher zu spezifizieren, weil wir lediglich einen Algorithmus für das Grundproblem entwerfen und uns nicht um die Teilmengenminimierung kümmern müssen. Zweitens ist er unter Umständen effizienter, da er weniger redundante Objekte speichert.

In der Tat hat sich empirisch gezeigt, dass diese Methodik zu einer deutlichen Effizienzsteigerung für verschiedene Probleme führt [Bl16]. Eine verbesserte Version des klassischen Dynamische-Programmierung-Algorithmus für grundiertes ASP ist mithilfe dieser Ideen implementiert worden [Fi17] und hat sich als erheblich schneller erwiesen als der Algorithmus aus [JPW09]. Unser Ergebnis formalisiert das gemeinsame Schema, das diesen Algorithmen zugrunde liegt. Auf diese Weise stellen wir einen formalen Rahmen bereit, der es ermöglicht, die erwähnten Optimierungen einfach auf andere Probleme zu übertragen. Dadurch machen wir die eindrucksvollen Effizienzsteigerungen, von denen in [Bl16, Fi17] berichtet worden ist, für Personen zugänglich, die an verwandten Problemen arbeiten. Das ist in erster Linie für Probleme auf der zweiten Stufe der Polynomiellen Hierarchie nützlich, da Teilmengenminimierung ein wiederkehrendes Thema vieler solcher Probleme ist.

### 3.2 Baumweitenerhaltende Klassen von nichtgrundiertem ASP

Wir definieren Klassen von nichtgrundierten ASP-Programmen, welche so grundiert werden können, dass die beschränkte Baumweite der Eingabe erhalten bleibt. Durch Einschränken der Syntax von nichtgrundiertem ASP definieren wir zwei Programmklassen, nämlich *bewachte* („guarded“) und *verknüpft bewachte* („connection-guarded“) Programme [Bl17]. Bewachte Programme garantieren, dass die Baumweite nach dem Grundieren immer dann klein ist, wenn die Baumweite der Eingabe klein ist. Wir beweisen diese Eigenschaft formal und zeigen, dass bewachte Programme trotz ihrer Einschränkungen weiterhin Probleme ausdrücken können, die vollständig für die zweite Stufe der Polynomiellen Hierarchie sind.

Verknüpft bewachte Programme sind sogar noch ausdrucksstärker als bewachte Programme. Wir zeigen, dass für verknüpft bewachte Programme die Baumweite nach dem Grundieren immer dann klein ist, wenn die Baumweite *und der maximale Knotengrad* der Eingabe (repräsentiert als Graph) klein ist.

Mit diesen Ergebnissen nähern wir uns dem Ziel, von der Empfindlichkeit, die moderne ASP-Solver augenscheinlich gegenüber Baumweite aufweisen, implizit zu profitieren, da sie uns Einblick in die Veränderung der Baumweite der Eingabe durch das Grundieren gewähren. So können wir, indem wir ein Programm in bewachtem ASP schreiben, sicher sein, dass moderne Grounder die Beschränkung der Baumweite nicht zerstören. Im Fall von verknüpft bewachtem ASP gilt das gleiche für die Kombination von Baumweite und maximalem Knotengrad.

**Beispiel 2.** Das ASP-Programm in Listing 1 kann verwendet werden, um zu entscheiden, ob eine gegebene Knotenmenge $S$ in einem gegebenen Graphen gesichert ist. Es rät eine Teilmenge $X$ von $S$ und verwendet sogenannte *schwache Constraints*, sodass die Kosten jedes Answer Sets exakt $|N[X] \cap S| - |N[X] \setminus S|$ betragen. (Wenn ein Programm schwache Constraints enthält, werden nur Answer Sets ausgegeben, die die Summe der verletzten schwachen Constraints minimieren.) Falls es eine Teilmenge von $S$ gibt, die weniger Nachbarn in $S$ als Nachbarn außerhalb von $S$ hat, dann gibt es ein Answer Set mit negativen Kosten. Wir können auf diese Weise entscheiden, ob $S$ gesichert ist, indem wir überprüfen, ob dieser minimale Wert negativ ist.

List. 1: Ein bewachtes ASP Programm, das prüft, ob eine gegebene Menge $S$ (deklariert mittels des Predikats `s`) in einem gegebenen Graph (deklariert mittels der Predikate `knoten` und `kante`) gesichert ist.

```
% Rate eine Teilmenge X von S.
x(S) | nx(S) :- s(S).
% Nachbarn von X sind "gut" wenn sie in S sind, ansonsten "boese".
nachbar(V) :- x(X), kante(X,V).
nachbar(X) :- x(X), knoten(X).
gut(V)     :- nachbar(V), s(V).
boese(V)   :- nachbar(V), knoten(V), not s(V).
% Hat X mehr boese Nachbarn als gute, so ist S nicht gesichert.
% Die folgenden schwachen Constraints stellen dies durch Summieren fest.
:~ knoten(V), gut(V).   [1,V]   % +1 fuer jeden guten Nachbarn.
:~ knoten(V), boese(V). [-1,V]  % -1 fuer jeden boesen Nachbarn.
```

Das Programm in Listing 1 ist bewacht. Man beachte, dass es alternativ auch ohne schwache Constraints möglich ist, zu überprüfen, ob eine Knotenmenge gesichert ist. Beispielsweise können wir die schwachen Constraints durch den „starken“ Constraint `:- #sum{ 1,G : gut(G); -1,B : boese(B)} >= 0` ersetzen. (Dieser Constraint enthält ein sogenanntes *Aggregat* – ein fortgeschrittenes ASP-Konstrukt, mit dem summiert werden kann.) Dieser neue Constraint ist jedoch nicht bewacht. Das bedeutet, dass das ursprüngliche Programm in Listing 1 im Allgemeinen zu Grundierungen von wesentlich kleinerer Baumweite führt und daher höhere Effizienz verspricht. △

In der Dissertation präsentieren wir ebenfalls eine Komplexitätsanalyse von Berechnungsproblemen, welche diesen Programmklassen entsprechen, und betrachten als Parameter die Baumweite der Eingabe, den maximalen Knotengrad der Eingabe, und eine Kombination dieser beiden Parameter. Die Ergebnisse dieser Analyse zeigen, dass ASP-Solving für jedes fixe bewachte ASP-Programm FPT ist, wenn wir die Baumweite der Eingabe als Parameter betrachten; darüber hinaus ist ASP-Solving für jedes fixe verknüpft bewachte Programm FPT, wenn der Parameter die Kombination aus Baumweite und maximalem Knotengrad ist. Diese Resultate sind nicht offensichtlich, da unsere ASP-Klassen schwache Constraints sowie Aggregate unterstützen, welche jeweils nicht von den FPT Algorithmen [JPW09, Fi17] für grundiertes ASP unterstützt werden. Desweiteren beweisen wir Schwereresultate, welche zeigen, dass für verknüpft bewachtes ASP *sowohl* Baumweite *als auch* maximaler Grad beschränkt sein müssen, um FPT zu erreichen. Zu diesem Zweck präsentieren wir eine verknüpft bewachte ASP-Kodierung eines Problems, das sogar für fixe Baumweite NP-schwer ist, und wir präsentieren eine bewachte Kodierung eines Problems, das sogar für fixen Knotengrad $\Sigma_2^P$-schwer ist.

Als Nebenprodukt dieser Untersuchungen erhalten wir *Metatheoreme* zum Beweisen von FPT-Resultaten. Mit anderen Worten: Unsere Ergebnisse zu bewachtem ASP ermöglichen es uns, zu beweisen, dass ein durch Baumweite parametrisiertes Problem FPT ist, indem wir dieses Problem einfach in bewachtem ASP ausdrücken. Wir vergleichen dieses Metatheorem mit dem verbreiteten Ansatz, FPT zu beweisen, indem man das Problem in

monadischer Prädikatenlogik zweiter Stufe ausdrückt und den wohlbekannten Satz von Courcelle anwendet. Auf ähnliche Weise können wir zeigen, dass ein durch Baumweite gemeinsam mit maximalem Knotengrad parametrisiertes Problem FPT ist, indem wir dieses Problem in verknüpft bewachtem ASP ausdrücken. Dieses Ergebnis ist ansprechend, da uns keine Metatheoreme bekannt sind, die es ermöglichen, FPT-Resultate für die Kombination von Baumweite und Grad als Parameter zu erhalten.

### 3.3 Allianzprobleme in Graphen

Wir führen eine Komplexitätsanalyse von Allianzproblemen in Graphen durch, sowohl im klassischen komplexitätstheoretischen Rahmen als auch parametrisiert durch Baumweite. Zunächst klären wir die Komplexität des SECURE SET Problems, indem wir zeigen, dass das Problem, sowie unterschiedliche Varianten, $\Sigma_2^P$-vollständig (und damit auf der zweiten Stufe der Polynomiellen Hierarchie) ist.

Dann widmen wir uns der Komplexität von SECURE SET und DEFENSIVE ALLIANCE wenn beide Probleme durch Baumweite parametrisiert sind. Wir verdeutlichen den Nutzen unserer ASP-Klassen als FPT-Klassifizierungswerkzeuge, indem wir einfache Kodierungen von Allianzproblemen präsentieren. Indem wir das NP-vollständige DEFENSIVE ALLIANCE Problem in verknüpft bewachtem ASP ausdrücken, erhalten wir auf einfache Weise das bereits bekannte Resultat, dass dieses Problem parametrisiert durch die Kombination von Baumweite und maximalem Knotengrad FPT ist. Von größerer Wichtigkeit ist jedoch, dass wir ein neues Resultat für das co-NP-vollständige Problem, zu entscheiden ob eine gegebene Knotenmenge in einem gegebenen Graph gesichert ist, erhalten: Wir zeigen, dass dieses Problem FPT für den Parameter Baumweite ist, indem wir das Problem in bewachtem ASP ausdrücken.

Wir liefern auch einige negative Resultate. So zeigen wir etwa, dass (unter weitverbreiteten komplexitätstheoretischen Annahmen) weder DEFENSIVE ALLIANCE noch SECURE SET FPT sind, wenn der Parameter die Baumweite ist. Diese Fragen sind seit der Vorstellung der Probleme in den Jahren 2002 bzw. 2007 unbeantwortet geblieben und wurden explizit als offene Probleme in [KO17] (bezüglich DEFENSIVE ALLIANCE) und [HD09] (bezüglich SECURE SET) genannt.

Trotz der parametrisierten Schwere von SECURE SET können wir zumindest ein leicht positives Ergebnis vermelden: Wir zeigen, dass das SECURE SET Problem immerhin in polynomieller Zeit gelöst werden kann, wenn die Instanzen beschränkte Baumweite haben, obwohl der Grad des Polynoms von der Baumweite abhängt.

## 4 Publikationen

Ein Großteil der Resultate wurde auf Konferenzen vorgestellt oder in Fachzeitschriften publiziert:

- Die Fortschritte in der Dynamischen Programmierung für Probleme mit Teilmengenminimierung wurden am AAAI-Workshop *Beyond NP 2016* vorgestellt und eine erweiterte Version in der Zeitschrift *Fundamenta Informaticae* veröffentlicht [Bl16].
- Die Klasse von verknüpft bewachten ASP-Programmen, wo Grundierungen unter Beibehaltung von beschränkter Baumweite möglich sind solange der maximale Knotengrad ebenfalls beschränkt ist, wurde auf der *IJCAI 2017* präsentiert [Bl17]. Der dort vorgestellte Artikel enthielt weder die gründliche Komplexitätsanalyse, die in der Dissertation durchgeführt wurde, noch die Arbeit über die Klasse der bewachten Programme, welche attraktiv sein kann, da hier der Knotengrad nicht beschränkt sein muss. Diese Zusätze sind derzeit für eine Konferenz unter Begutachtung. Ein Zeitschriftenartikel ist in Planung.
- Das $\Sigma_2^P$-Vollständigkeitsresultat für das SECURE SET Problem wurde auf der Konferenz *WG 2015* vorgestellt [BW16]. Eine erweiterte Version dieses Artikels, welche zusätzlich die Ergebnisse zur parametrisierten Komplexität des Problems enthält, wurde in der Zeitschrift *Algorithmica* veröffentlicht [BW17]. Die vorliegende Dissertation enthält zusätzlich die parametrisierte Komplexitätsanalyse des DEFENSIVE ALLIANCE Problems, wozu ein Zeitschriftenartikel derzeit unter Begutachtung ist.

Die Forschung am Dissertationsthema hat darüber hinaus zu zahlreichen Nebenprodukten geführt, die unter anderem auf den Konferenzen *IJCAI 2016* (zwei Artikel), *ECAI 2016*, *FoIKS 2016*, *COMMA 2016* und *JELIA 2014* vorgestellt wurden. Außerdem sind zwei Artikel im *Journal of Logic and Computation (JLC)* erschienen.

## Literaturverzeichnis

[Ab15] Abseher, Michael; Bliem, Bernhard; Charwat, Günther; Dusberger, Frederico; Woltran, Stefan: Computing Secure Sets in Graphs Using Answer Set Programming. J. Logic Comput., 2015. Zur Publikation angenommen.

[ALS91] Arnborg, Stefan; Lagergren, Jens; Seese, Detlef: Easy Problems for Tree-Decomposable Graphs. Journal of Algorithms, 12(2):308–340, 1991.

[BDH07] Brigham, Robert C.; Dutton, Ronald D.; Hedetniemi, Stephen T.: Security in Graphs. Discrete Appl. Math., 155(13):1708–1714, 2007.

[BET11] Brewka, Gerhard; Eiter, Thomas; Truszczyński, Mirosław: Answer Set Programming at a Glance. Communications of the ACM, 54(12):92–103, 2011.

[Bl16] Bliem, Bernhard; Charwat, Günther; Hecher, Markus; Woltran, Stefan: D-FLATˆ2: Subset Minimization in Dynamic Programming on Tree Decompositions Made Easy. Fund. Inform., 147(1):27–61, 2016.

[Bl17] Bliem, Bernhard; Moldovan, Marius; Morak, Michael; Woltran, Stefan: The Impact of Treewidth on ASP Grounding and Solving. In (Sierra, Carles, Hrsg.): Proceedings of IJCAI 2017. AAAI Press, S. 852–858, 2017.

[Bo93] Bodlaender, Hans L.: A Tourist Guide through Treewidth. Acta Cybernet., 11(1-2):1–21, 1993.

[BW16] Bliem, Bernhard; Woltran, Stefan: Complexity of Secure Sets. In (Mayr, Ernst W., Hrsg.): Revised Papers of WG 2015. Jgg. 9224 in LNCS. Springer, S. 64–77, 2016.

[BW17] Bliem, Bernhard; Woltran, Stefan: Complexity of Secure Sets. Algorithmica, 2017. Im Druck.

[Fi17] Fichte, Johannes Klaus; Hecher, Markus; Morak, Michael; Woltran, Stefan: Answer Set Solving with Bounded Treewidth Revisited. In (Balduccini, Marcello; Janhunen, Tomi, Hrsg.): Proceedings of LPNMR 2017. Jgg. 10377 in LNCS. Springer, S. 132–145, 2017.

[Fl02] Flake, Gary William; Lawrence, Steve; Giles, C. Lee; Coetzee, Frans: Self-Organization and Identification of Web Communities. IEEE Computer, 35(3):66–71, 2002.

[FRV14] Fernau, Henning; Rodríguez-Velázquez, Juan A.: A Survey on Alliances and Related Parameters in Graphs. Electron. J. Graph Theory Appl. (EJGTA), 2(1):70–86, 2014.

[HD09] Ho, Yiu Yu; Dutton, Ronald D.: Rooted Secure Sets of Trees. AKCE Int. J. Graphs Comb., 6(3):373–392, 2009.

[HHH03] Haynes, Teresa W.; Hedetniemi, Stephen T.; Henning, Michael A.: Global Defensive Alliances in Graphs. Electron. J. Combin., 10, 2003.

[JPW09] Jakl, Michael; Pichler, Reinhard; Woltran, Stefan: Answer-Set Programming with Bounded Treewidth. In (Boutilier, Craig, Hrsg.): Proceedings of IJCAI 2009. AAAI Press, S. 816–822, 2009.

[KO17] Kiyomi, Masashi; Otachi, Yota: Alliances in Graphs of Bounded Clique-Width. Discrete Appl. Math., 223:91–97, 2017.

[Mo10] Morak, Michael; Pichler, Reinhard; Rümmele, Stefan; Woltran, Stefan: A Dynamic-Programming Based ASP-Solver. In (Janhunen, Tomi; Niemelä, Ilkka, Hrsg.): Proceedings of JELIA 2010. Jgg. 6341 in LNCS. Springer, S. 369–372, 2010.

**Bernhard Bliem** wurde 1988 geboren und studierte an der TU Wien Informatik (Studiengänge *Software & Information Engineering* und *Computational Intelligence*) sowie Philosophie an der Universität Wien. Seine Diplomarbeit wurde mit dem *Distinguished-Young-Alumnus*-Preis der Fakultät für Informatik der TU Wien ausgezeichnet. Das Doktorat in Informatik absolvierte er unter der Betreuung von Prof. Stefan Woltran an der TU Wien. Nach Abschluss seines Doktorats begann Bernhard Bliem als Postdoc an der Universität Helsinki zu forschen. Seine Forschungsinteressen umfassen verschiedene Themen aus der Künstlichen Intelligenz wie etwa Answer Set Programming, SAT Solving, Logikprogrammierung und Wissensrepräsentation, sowie Algorithmik und Komplexitätstheorie.

# Multi-modale 3D-Kartierung – Kombination von 3D-Punktwolken mit Thermo- und Farbinformationen[1]

Dorit Borrmann[2]

**Abstract:** Man stelle sich eine Technologie vor, die automatisch ein vollständiges 3D-Thermographiemodell einer Umgebung generiert und Temperaturspitzen darin erkennt. Die Analyse einer Umgebung bezüglich Energieeffizienz oder zur Überwachung wichtiger Infrastruktur anhand von Thermalbildern ist zeitaufwändig und nur durch Erfahrung und Expertise möglich. Die hier vorgestellte Arbeit [Bo17] schlägt ein Robotersystem vor, das durch Kombination von Thermographie mit terrestrischem Laserscanning ein vollständiges 3D-Modell der Umgebung mit Farb- und Temperaturinformation erstellt. Die ergänzende Farbkamera vereinfacht die Interpretation der Daten und eröffnet weitere Anwendungsfelder. Die an unterschiedlichen Positionen aufgenommenen Daten aller Sensoren werden durch Kalibrierung und Scanmatching in einem gemeinsamen Bezugssystem zusammengefügt. Die Arbeit beschreibt und evaluiert die hierzu benötigten Verfahren und zeigt Methoden zur Weiterverarbeitung der Daten auf.

Ein vollständiges multi-modales 3D-Modell enthält alle relevanten geometrischen Informationen der aufgenommenen Szene und ermöglicht einem Experten, diese standortunabhängig zu analysieren. Diese Technologie ebnet den Weg für die automatische Erkennung relevanter Bereiche und für die Analyse des Wärmeflusses und vereinfacht somit die Lokalisierung und Identifikation von Wärmelecks für den Experten. Das vorgestellte modulare Konzept ist weder auf den Anwendungsfall Energieeffizienz beschränkt noch auf die Verwendung einer mobilen Plattform angewiesen. Es ist beispielsweise auch in Feldern wie der Archäologie und Geologie einsetzbar und kann durch zusätzliche Sensoren erweitert werden.

## 1 Einführung

Fortschritte in Technologie und Forschung führen zur vermehrten Bemühungen Umgebungen am Computer zu analysieren. Die Kombination unterschiedlicher Sensordaten erzeugt ein vollständiges Umgebungsbild. Ein Foto ist in der Regel vom Benutzer schnell zu erfassen hat aber ein eingeschränktes Sichtfeld. Auch in Panoramabildern, die diesen Mangel kompensieren, fehlen die geometrischen Informationen weitestgehend. Die Rekonstruktion der Geometrie einer Szene durch die Verwendung von Bildmerkmalen scheitert in merkmalsarmen Umgebungen und ist nicht skaliergenaus. Bei Aufgaben in der Bauindustrie, der Archäologie oder der Geologie, die geometrisch korrekte Messwerte verlangen, haben sich in den letzten Jahren 3D-Laserscanner durchgesetzt. Neben Entfernungsmesswerten liefern diese heutzutage häufig auch die Intensität des reflektierten Lichtstrahls. Wenn das Einfärben der Punktwolke mit diesen Intensitätswerten auch die Interpretation einer Szene erleichtert, so wirkt es doch für einen ungeübten Benutzer schwer erkenntlich. In Zeiten reger Diskussion über Energieeffizienz ist Thermografie

[1] Englischer Originaltitel der Dissertation: “Multi-modal 3D mapping – Combining 3D point clouds with thermal and color information”

[2] Lehrstuhl für Robotik und Telematik, Universitiät Würzburg, borrmann@informatik.uni-wuerzburg.de

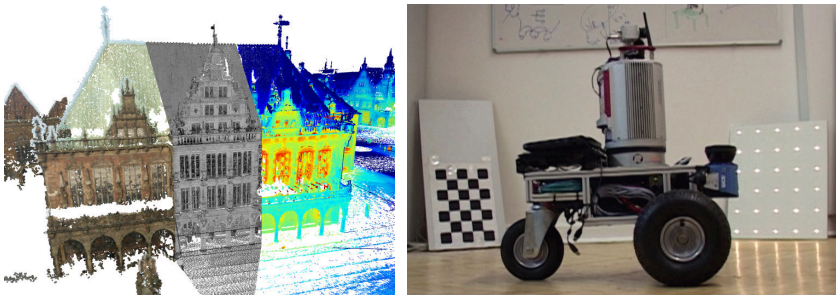


Abb. 1: Links: Skizze eines multi-modalen Modells des Bremer Rathauses, eine Punktwolke mit Temperatur- und Farbinformationen. Rechts: Der Roboter Irma3D vor den Kalibriermustern.

sehr gefragt. Sie kommt auch für die Sicherung wichtiger Infrastruktur, wie Energieversorgung und Temperaturregulierungssystemen, zum Einsatz. Mit Infrarotkameras lässt sich das Wärmebild eines Gebäudes dokumentieren. Für die Interpretation und genaue Lokalisierung auffälliger Bereiche wäre jedoch ein komplettes 3D-Modell mit Thermaldaten und Echtfarben wünschenswert. Bisher wird aufwendig per Hand modelliert. Erschwerend kommt bei Außenaufnahmen bereits der Einfluss diffuser Sonneneinstrahlung an einem bewölkten Tag hinzu. Thermografie bei Nacht verhindert jedoch die zeitgleiche Aufnahme von Fotos und Thermalbildern.

Die vorgestellte Arbeit löst dieses Problem mit einem Robotersystem, das durch Kombination von Thermographie mit terrestrischem Laserscanning ein vollständiges 3D Modell der Umgebung mit Farb- und Temperaturinformationen erstellt (vgl. Abb. 1). Die ergänzende Farbkamera vereinfacht die Dateninterpretation und eröffnet weitere Anwendungsfelder. Kalibrierung und Scanmatching fügen die an unterschiedlichen Positionen aufgenommenen Daten aller Sensoren in einem gemeinsamen Beszugssystem zusammen. Der erste Teil der Arbeit behandelt 3D-Punktwolkenverarbeitung mit Schwerpunkt auf effizientem Punktzugriff, Erkennung planarer Strukturen und Registrierung mehrerer Punktwolken in einem Koordinatensystem. Der zweite Teil beschreibt die autonome Erkundung und Datenakquise mit einem mobilen Roboter, mit dem Ziel, die bisher nicht erfassten Bereiche im 3D-Raum zu minimieren. Die Kombination von Farbbildern, Thermalbildern und Punktwolken durch Kalibrierung wird ausgearbeitet. Den abschließenden Teil stellen Anwendungsszenarien für die gesammelten Daten dar, darunter Methoden zur Erkennung der Innenraumstruktur für die Rekonstruktion von Gebäuden und der anschließenden Klassifizierung von Fenstern. Ein System zur Rückprojektion der Thermalinformation in die Umgebung wird ebenso vorgestellt wie Methoden zur Verbesserung der Farbinformationen und zum Zusammenfügen separat aufgenommener Punktwolken und Fotoreihen.

### 1.1 Wissenschaftlicher Beitrag

Die Entwicklung der 3D-Messtechnik in den letzten Jahren und die damit verbundene Kostenreduktion hat die 3D-Punktwolkenverarbeitung in den Fokus vieler Anwendungsbereiche gerückt, darunter Archäologie und Denkmalpflege, Geologie sowie die Bau- und

Unterhaltungsindustrie. Mit der Verfügbarkeit schneller Messgeräte kommt die Notwendigkeit Daten effizient zu verarbeiten. Die vorgestellte Arbeit ist das Ergebnis von Forschung im Bereich der Robotik mit dem Schwerpunkt des Laserscannings und der 3D-Punktwolkenverarbeitung. Sie stellt das Bestreben dar, 3D-Punktwolken automatisch mit anderen Modalitäten, wie Farb- und Thermalbildern, zu kombinieren und diese so aufzubereiten, dass sie in diversen Anwendungen verwendet werden können. Der wissenschaftliche Beitrag lässt sich in den Kategorien 3D Datenstrukturen, Ebenenerkennung, Punktwolken-Registrierung, 3D Exploration und thermale Modellierung zusammenfassen. Die Grundlagenforschung in der 3D-Punktwolkenverarbeitung in diesen Kategorien kommt in diversen Anwendungen in unterschiedlichen Bereichen zum Einsatz.

## 2 Datenakquise

Der mobile Roboter Irma3D (Abk. Intelligent robot for mapping applications in 3D, siehe Abb.1) wurde speziell für die 3D-Kartierung entwickelt. Die beiden Antriebsräder verfügen über Radencoder zur Berechnung der Odometrie, die zur Bewegungsschätzung mit einer inertialen Messeinheit (IMU) fusioniert wird. Ein 2D-Laserscanner dient zur Hindernisvermeidung und Navigation. Für die Datenakquise verwendet Irma3D einen Riegl VZ-400 3D-Laserscanner. Auf diesem werden wahlweise eine Optris PI Imager Thermokamera, eine Webcam oder eine Spiegelreflexkamera angebracht. Die Roboterkontrollarchitektur ist in ROS (Abk. Robot Operating System) realisiert.

Irma3D erkundet eine Umgebung autonom mit einem zweigeteilten Pfadplannungsalgorithmus. An der Startposition wird ein 3D-Laserscan mit $360^\circ \times 100^\circ$ Öffnungswinkel aufgenommen und mehrere Fotos, um den horizontalen Bildwinkel von $360^\circ$ abzudecken. In einer Linienkarte, generiert aus einem horizontalen Schnitt der Punktwolke auf Höhe des Roboters, markieren Sprungkanten Bereiche zwischen explorierten und nicht explorierten Bereichen. Der Roboter fährt für die nächste Aufnahme eine Position vor der Sprungkante an, bei der die meisten neuen Informationen zu erwarten sind. Sobald ein Raum, d.h. ein abgeschlossener Bereich, erkannt wird, wechselt die Pfadplanung in den 3D-Modus, bei dem die Anzahl der ungesehenen Voxel in einer 3D-Karte minimiert wird. Anschließend wird die Pfadplanung in 2D fortgesetzt. Abb. 2 zeigt die 2D-Karte und die Scanpositionen an einer beispielhaften Büroumgebung. Der Mehrwert durch die 3D-Exploration wird in [Bo14] gezeigt.

## 3 3D-Punktwolkenverarbeitung

Moderne Laserscanner erfassen die Umgebung präzise mit einer Datenrate im sechs- bis siebenstelligen Bereich pro Sekunde. Die Rohdatenverarbeitung ist somit schon aufgrund der Menge und dem damit verbundenen Zeit- und Speicheraufwand problematisch. Aufgabenabhängig muss ein schneller Zugriff bei geringem Speicheraufwand gewährleistet werden. In [Bo17] kommen zu diesem Zwecke Baumstrukturen, welche die räumliche Struktur der Daten nutzen, namentlich Octrees und $k$D-Bäume, und den Aufnahmeprozess widerspiegelnde Panoramabilder zum Einsatz.

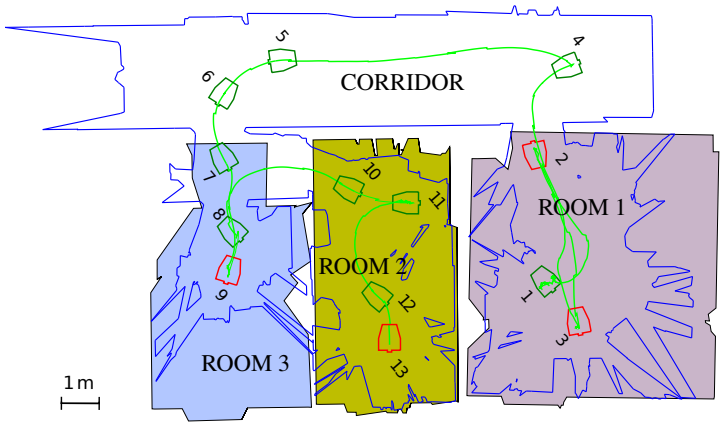



Abb. 2: Roboterpfad und Scanpositionen während der Exploration. Grün markiert sind die Posen aus der 2D-Pfadplanung, während rote Posen im 3D-Modus angefahren wurden. (Video eines multimodalen Modells der Versuchsumgebung: `http://youtu.be/qoQlP8F0zg0`)

### 3.1 Ebenenerkennung

Die Verwendung primitiver Formen reduziert die Verarbeitungszeit weiter. In künstlich geschaffenen Umgebungen sind Ebenen dominant. Drei grundlegende Ansätze detektieren diese in Punktwolken. RANSAC (engl. Random Sample Consensus) Algorithmen sind generelle randomisierte Verfahren, um für beobachtete Daten ein parametrisiertes Modell zu finden. Hierzu wird mehrfach eine minimale Menge an Daten, die das Modell definieren, ausgewählt und überprüft, wie gut dieses Modell die Daten repräsentiert. Die Hough Transformation sucht im Parameterraum die Ebene, auf der die meisten Punkte liegen und transformiert dazu die Punkte vom kartesichen Raum in den Raum der Ebenenparameter. Bei Region Growing Ansätzen wird die Ebene sukzessive durch Nachbarpunkte vergrößert. In der vorgestellten Arbeit werden Algorithmen von jedem Typ für die Erkennung eines Kalibriermusters und die Bestimmung der Gebäudestruktur analysiert. Während die Hough Transformation die dominanten Ebenen zuverlässiger findet als Region Growing Ansätze, verhält es sich für die Erkennung eines klar definierten Musters andersherum.

### 3.2 3D-Kartierung

Um ein komplettes verschattungsfreies Umgebungsmodell zu erhalten, müssen Laserscans an unterschiedlichen Positionen aufgenommen und registriert, d.h. in ein gemeinsames Koordinatensystem gebracht werden. Das Registrierungsverfahren besteht aus zwei Komponenten, einer paarweisen Registrierung und einer globalen Optimierung [Bo08].

**Registrierung mittels 3D-Scanmatching** Für die paarweise Registrierung verwenden wir den bekannten ICP-Algorithmus (Abk. Iterative Closest Point) [BM92]. Der ICP ist in

der Lage eine lokale Verbesserung der initialen Transformationen einer Sequenz von 3D-Scans zu erzeugen. Sei $M$ die Modellmenge der 3D-Punkte des Scans mit fester Transformation und $D$ die Datenmenge von Punkten des zu matchenden Scans, dann berechnet der ICP die Transformation von $D$ basierend auf Punktkorrespondenzen zwischen den beiden Punktmengen. Iterativ werden die metrisch nächstgelegenen Punkte aus beiden Punktmengen als korrespondierend ausgewählt, sofern ihr Abstand unter einem Schwellwert liegt. Berechnet wird nun die Transformation $(R,t)$, welche die Fehlerfunktion

$$E_{\text{ICP}}(R,t) = \frac{1}{N}\sum_{i-1}^{N} ||m_i - (Rd_i + t)||^2 \tag{1}$$

minimiert. $N$ ist die Anzahl der Punktpaare in $M$ und $D$. Unter Annahme korrekter Punktkorrespondenzen in der letzten Iteration lässt sich in geschlossener Form die Transformation berechnen, welche die Fehlerfunktion minimiert.Um Umgebungen vollständig zu digitalisieren, müssen mehrere 3D-Scans registriert werden. Diese 3D-Scans seien in einer Sequenz von $n+1$ 3D-Scanposen $V_0,\ldots,V_n$ gespeichert. Die einfachste Methode zum Registrieren mehrere 3D-Scans ist das so genannte paarweise Scanmatching. Hierbei wird die Modelmenge $M$ von dem 3D-Scan der Pose $V_{j-1}$ und die Datenmenge $D$ von dem 3D-Scan an Pose $V_j$ für alle $j$ in $[1,n]$ gebildet. Alternativ kann man als Modelmenge auch die Vereinigung aller bereits registrierter Scans verwenden. Diese Methode wird als Metascan-Matching bezeichnet. Zur Bestimmung der initialen Transformation eignen sich die Odometrie des Roboters, GNSS-Messungen, 2D-Kartierung oder merkmalsbasierte Registrierung. Eine Untersuchung dazu findet sich in [Bo17].

**Global konsistentes Scanmatching** Sowohl paarweises als auch Metascan-Matching korrigieren die Poseschätzungen für die einzelnen 3D-Scans. Dennoch summieren sich Registrierungsfehler auf. SLAM-Algorithmen aus der Robotik verwenden die Methode des Schleifenschließens, um diese Fehler zu begrenzen. Falls die Differenz der Poseschätzungen $V_j = (x_j, y_j, z_j, \theta_{xj}, \theta_{yj}, \theta_{zj})$ und $V_k = (x_k, y_k, z_k, \theta_{xk}, \theta_{yk}, \theta_{zk})$ zweier 3D-Scans nach paarweisem bzw. Metascan-Matching unterhalb eines Schwellwertes liegen, nehmen wir an, dass sich diese Scans matchen lassen. Dem korrespondierenden Graphen, der anfänglich mit der Scanposensequenz $((V_0,V_1),\ (V_1,V_2),\ldots,(V_{n-1},V_n))$ initialisiert wird, wird die Kante $(V_j,V_k)$ hinzugefügt.

Nachdem die 3D-Scans mit paarweisem bzw. Metascan-Matching registriert worden sind und der Graph erzeugt wurde, wenden wir eine globale Relaxation an. Gegeben sei ein Netz mit $n+1$ Knoten $X_0,\ldots,X_n$, das die Posen $V_0,\ldots,V_n$, und die gerichteten Kanten $D_{j\to k}$ repräsentiert. Das Ziel ist nun, eine global konsistente Karte zu erzeugen, also die Posen $V_0,\ldots,V_n$ so zu schätzen, dass alle 3D-Scans konsistent registriert werden. Die Fehlerfunktion wird erweitert, so dass alle Kanten des Graphen berücksichtigt werden

$$E_{\text{opt}} = \sum_{j\to k}\sum_{i=1}^{N_{j\to k}} ||R_j m_i + t_j - (R_k d_i + t_k)||^2. \tag{2}$$

Die Fehlerfunktion beinhaltet die Punktkorrespondenzen für alle Scans, die durch eine Kante verbunden sind. Gesucht sind die Transformationen, die das Netz von Korrespon-

denzen minimieren. Zur Minimierung der Fehlerfunktion muss diese linearisiert werden. Für 3D-Scans und 6D-Posen ist dies in [Bo08] beschrieben. Im vollständigen Algorithmus zur schrittweisen Optimierung der Scanposen wird jeder Scan zuerst an seinen Vorgänger registriert. Sobald eine Schleife erkannt wird, d.h., der Abstand zu einem bereits registrierten Scan ist klein genug, wird automatisch ein Graph erstellt, der die Nachbarschaftsbeziehungen der Scans darstellt und das zugehörige Gleichungssystem minimiert. Dies wird iterativ wiederholt, bis die Veränderung der Posen klein ist.

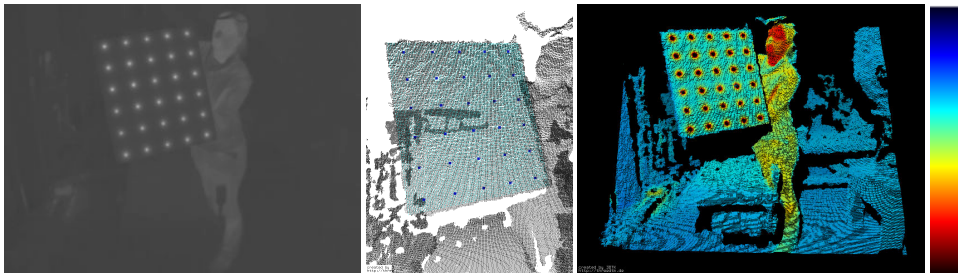
## 4 Erstellung der 3D-Thermomodelle

Bei der gleichzeitigen Aufnahme mit mehreren Sensoren sieht ein jeder Sensor die Welt in seinem eigenen Koordinatensystem. Die Zuordnung zueinander wird vereinfacht, wenn die Sensoren zueinander kalibriert sind, d.h., ihre relative Positionierung bekannt ist. Um dies zu erreichen, benötigt man ein Kalibrierungsmuster, das in den Daten beider Sensoren eindeutig erkannt wird. Für die Kalibrierung von Kameras wird üblicherweise ein Schachbrettmuster verwendet, da die Ecken des Musters gut in Fotos identifizierbar sind. Für Infrarotkameras eignet sich dieses Muster schlechter, da die Kanten selbst nach vorheriger Bestrahlung in der Aufnahme sehr verschwommen erscheinen. Als Alternative bieten sich eindeutig identifizierbare Wärmequellen an. Abb. 1 zeigt ein Kalibrierungsmuster für Infrarotkameras, bei dem 30 kleine Glühlampen gleichmäßig in einem Raster auf einem Brett angeordnet sind. Diese erscheinen deutlich in den Aufnahmen. Um die Kalibrierungsmuster auch in den Laserscans zu lokalisieren, werden sie auf einem Brett angebracht. Für Farbkameras kommt das gleiche Verfahren mit einem Schachbrettmuster zum Einsatz (vgl. Abb. 1). Durch Evaluation verschiedener Ebenenerkennungsverfahren wurde ein Verfahren entwickelt, das mit hoher Wahrscheinlichkeit automatisch, robust und präzise das Kalibriermuster in Punktwolken erkennt. Gegeben die Spezifikation des Kalibrierungsmusters detektiert Algorithmus 1 das Brett in einer Punktwolke.

**Algorithmus 1** Kalibriermustererkennung in einem Laserscan.

**Require:** Punktwolke, Spezifikation des Kalibriermusters
1: Entferne alle Punkte außerhalb des Erwartungsbereichs des Brettes.
2: Finde die Ebene, die das Brett enthält
3: Projiziere ein generiertes Modell des Brettes in den Mittelpunkt der detektierten Ebene.
4: Verwende den ICP-Algorithmus um das Modell an die Datenpunkte anzupassen.
5: **if** Jeder Punkt des Ebenenmodells hat eine Korrespondenz in der Punktwolke **then**
6: **return** Position der Glühlampen anhand des ICP Ergebnisses
7: **end if**

Die Erstellung eines 3D-Thermalmodells ist in Abb. 3 dargestellt. Mit den Glühlampen in den Thermogrammen wird zuerst die Wärmebildkamera intrinsich und dann entsprechend Algiorithmus 1 extrinsich zum 3D-Sensor kalibriert. Um die Genauigkeit zu erhöhen, insbesondere aufgrund der geringen Auflösung der Infrarotkamera, werden mehrere Datenpaare aus Thermalbild und Punktwolke augenommen und durch Gradientenabstieg die optimale Transformation bestimmt. Mit der so bestimmten Transformation werden die 3D-Punkte auf das Bild projiziert und die Temperaturwerte den Punkten zugeordnet.



(a) (b) (c) (d)

Abb. 3: Erstellung einer thermalen Punktwolke. (a) Thermogramm (weiß $\hat{=}$ 0 °C, schwarz $\hat{=}$ 100 °C). (b) Das Modell (cyan) wird bei der Kalibrierung an die Daten registriert um die Position der Glühlampen (blau) zu bestimmen. (c) Punkte nach der Skala aus (d) von 10 °C to 40 °C eingefärbt.

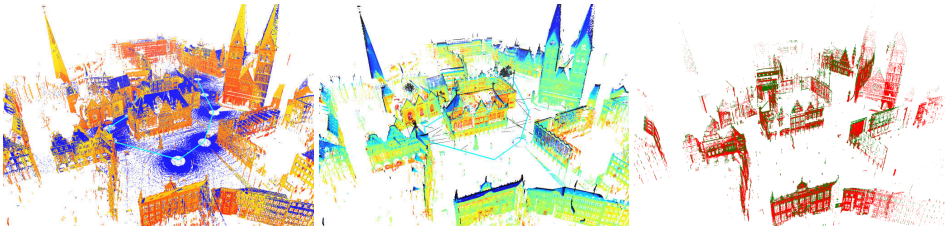
Abb. 4: Ein Scan aus dem Kaisersaal der Würzburger Residenz eingefärbt aus Fotos ohne (links) und mit (rechts) Korrektur mittels Raytracing. (Video: `http://youtu.be/jKVxlLvu7Pk`).

### 4.1 Ergebnisse

Das Verfahren zur Kombination von 3D-Punktwolken mit Temperatur- und Farbinformationen wurde an zahlreichen Datensätzen visuell überprüft. Für eine Evaluation der einzelnen Schritte des Kalibrierverfahrens sei auf [Bo17] verwiesen. Im folgenden wird an einigen Beispielen der Mehrwert durch die 3D-Geometrie dargestellt.

Abb. 4 zeigt ein 3D-Modell, das Irma3D im Kaisersaal der Würzburger Residenz aufgenommen hat. Aufgrund des Versatzes zwischen Laserscanner und Kamera ist die Einfärbung durch Projektion der Punkte auf das Kamerabild im Bereich von Verschattungen fehlerhaft. Darum wurde ein Raytracing-Verfahren entwickelt, mit dem für jeden Punkt der Punktwolke überprüft wird, ob dieser von der Kamera aus sichtbar ist, oder durch andere Punkte verdeckt wird. Die Punkte sind in einem $k$D-Baum organisiert, was neben schnellem Verwerfen nicht zu betrachtender Punkte die parallele Verarbeitung erlaubt. Im Kaisersaal erkennt man die Korrektur deutlich hinter den Kronleuchtern und den Stativen.

Die bei der Thermographie gemessene abgegebene Strahlung einer Oberfläche wird maßgeblich durch die Emission $\varepsilon$, die Reflektion $\rho$, und die Transmission $\tau$ beeinflusst. Für



(a) Sichtwinkel (b) Bereinigt (c) Temperaturabweichungen

Abb. 5: Der Bremer Marktplatz (Video: `http://youtu.be/TPoCebERysc`) (a) Punkte eingefärbt anhand des Winkels zwischen Normale und Sichtvektor. Blaue Punkte haben einen Winkel größer als 60°. (b) Die Szene ohne die Punkte mit einem Winkel größer als 60° eingefärbt anhand der Temperaturwerte. (c) Punkte mit Nachbarn im Abstand von 1 cm. Grün markierte Punkte haben eine Temperaturabweichung von $> 1\,\mathrm{K}$.

aussagekräftige Messwerte muss die Reflektion der Umgebungsstrahlung durch günstige Wahl der Aufnahmewinkel gering gehalten werden. Da diese bei der Aufnahme schwer zu bestimmen sind, werden nachträglich jene Punkte gelöscht, bei denen die Winkelbegrenzung von 60° nicht eingehalten wurde. Im Octrees werden die nächsten Nachbarn eines jeden Punktes und mit ihnen die Normale bestimmt. Ist der Winkel zwischen Messrichtung und Normale zu groß, gilt die Messung als unzuverlässig. Abb. 5 zeigt das Ergebnis am Beispiel des Bremer Marktplatzes. Rund um das Rathaus hat Irma3D an 13 Positionen Laserscans und Wärmebilder aufgenommen. Von den 81.398.810 Punkten erfüllen 62.272.650 Punkte das Winkelkriterium. Insbesondere Dächer und Boden, aber auch entfernte Wände, erfasst der bodennahe Roboter unzureichend.

Die überlappenden Aufnahmen erfassen große Bereiche mehrfach. Dies dient ebenfalls der Überprüfung der Messungen. Bestimmt man für jeden Punkt den jeweils nächsten Punkt aus allen anderen Scans, so lassen sich alle Punkte mit einer Abweichung entfernen, wie in Abb. 5(c) exemplarisch für Nachbarn mit einem Maximalabstand von einem Zentimeter und einer Temperaturabweichung von 1 Kelvin dargestellt.

## 5 Anwendungen

Die in der Arbeit entwickelten Methoden bilden die Grundlage für zahlreiche weitere Anwendungen die hier kurz zusammengefasst werden. Basierend auf der Ebenenerkennung werden die Strukturelemente von Innenräumen bestimmt. Dies ermöglicht die Rekonstruktion von verdeckten Bereichen (vgl. Abb. 6(a)) oder in einem Thermalmodell die Klassifizierung von Fenstern als offen oder geschlossen (vgl. Abb. 6(b) und 6(c)). Wird der Laserscanner durch eine 3D-Kamera ersetzt und das System mit einem Projektor erweitert, so können die Thermaldaten direkt wieder in die Szene zurück projiziert werden. Beispielsweise in der Fertigung werden dadurch die entstanden Wärmeverteilungen direkt sichtbar gemacht (vgl. Abb. 6(f)). Bei der zeitversetzen Aufnahme von Farbfotos verändern die wechselnden Lichtverhältnisse die farbliche Darstellung einer Scene. In einem Modell entstehen dadurch deutlich sichtbare Übergänge. Durch Ausnutzung der räumlichen

Struktur des Modells kann ein Gleichungssystem aufgestellt werden, um die Bilder radiometrisch zu korrigieren (vgl. Abb. 6(d) und 6(e)). Um den Einfluss der Sonnenstrahlung zu vermeiden, wird Thermographie nachts durchgeführt. Dies bedeutet, dass Farbbilder nicht zeitgleich aufgenommen werden können. Soll das Modell nun trotzdem mit Farbwerten bereichert werden, muss die Transformation anders bestimmt werden. Eine Möglichkeit ist es, ein 3D-Modell durch Photogrammetrie zu erstellen, dieses an das Lasermodell zu registrieren und dadurch die Transformation der Fotos zu berechnen. Dies erhöht die Dichte insbesondere in strukturarmen Regionen (vgl. Abb. 6(g) und 6(h)).

## 6 Zusammenfassung

Die vorgestellte Arbeit untersucht die Kombination von 3D-Punktwolken mit Farb- und Thermobildern mit dem Schwerpunkt auf der Kokalibrierung zwischen Laserscanner und Wärmebildkamera. Das entwickelte Glühlampenmuster wird sowohl im Laserscan als auch im Thermobild zuverlässig erkannt. Da das Verfahren direkt auf Farbbilder übertragbar ist, ist es in Bereichen wie der Geologie und der Archäologie anwendbar. Die Ebenenerkennungsmethoden, die bei der Erkennung des Kalibriermusters zum Einsatz kommen, sind auch für weitere Verarbeitund des fertigen Modells hilfreich. Global agierende Ebenendetektionsverfahren erweisen sich als vorteilhaft, wenn die Hauptstruktur von Gebäuden erkannt werden soll, während Region Growing Ansätze besser kleinere Strukturen detektieren. Zusätzlich zur Kalibrierung, die Daten unterschiedlicher Sensoren kombiniert, vereint das Scanmatching Daten unterschiedlicher Positionen. Die globale Optimierung hilft dabei, ein global konsistentes Modell zu erzeugen. Um die enorme Datenmenge verarbeiten zu können, kommen effiziente Implementierungen von Octree, *k*D-Baum und Panoramen zum Einsatz. Das 3D-Explorationsverfahren zur autonomen Erstellung eines 3D-Modells erreicht eine bessere Abdeckung als das 2D-Verfahren. Die entwickelte Technologie wurde erfolgreich in zahlreichen Umgebungen getestet und ist in vielerlei Hinsicht erweiterbar. Zusätzlich zu den vorgestellten Anwendungsszenarien sind noch viele weitere denkbar. Zu den nächsten Zielen gehören die Beschleunigung der Aufnahme durch mobiles Laserscanning, die Erweiterung der Fensterklassifizierung auf andere Objekte und die Erstellung von Oberflächenmodellen aus den Punktwolken.

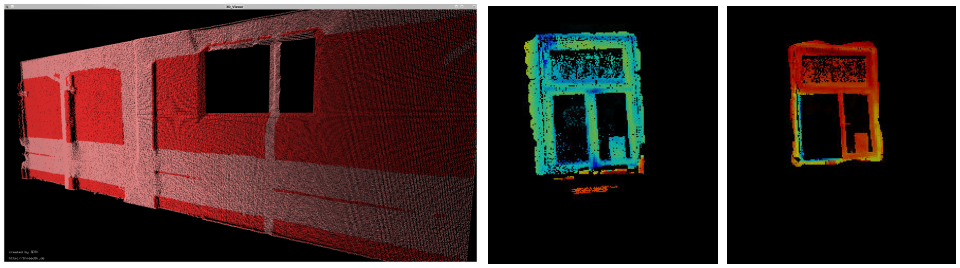
## Literaturverzeichnis

[BM92] Besl, P.; McKay, N.: A Method for Registration of 3–D Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(2):239 – 256, February 1992.

[Bo08] Borrmann, D.; Elseberg, J.; Lingemann, K.; Nüchter, A.; Hertzberg, J.: Globally consistent 3D mapping with scan matching. Journal of Robotics and Autonomous Systems (JRAS), 56(2):130–142, February 2008.

[Bo14] Borrmann, D.; Nüchter, A.; Đakulović, M.; Maurović, I.; Petrović, I.; Osmanković, D.; Velagić, J.: A mobile robot based system for fully automated thermal 3D mapping. Advanced Engineering Informatics, 28(4):425–440, October 2014.

[Bo17] Borrmann, D.: Multi-modal 3D mapping – Combining 3D point clouds with thermal and color information. Doctoral thesis, Universität Würzburg, 2017.



(a) rekonstruierte Wand (b) geschlossen Fenster (c) offenes Fenster

(d) ohne Korrektion (e) radiometrisch korrigiert

(f) projizierte Temperaturwerte (g) Photogrammetrie (h) eingefärbter Laserscan

Abb. 6: Anwendungsbeispiele.

**Dorit Borrmann** wurde am 21. März 1984 in Haselünne geboren. Seit Juli 2013 ist sie als wissenschaftliche Mitarbeiterin am Institut für Informatik - Robotik und Telematik der Julius-Maximilians-Universität Würzburg beschäftigt, wo sie im Dezember 2017 ihre Promotion abschloss. Zuvor war sie an der Jacobs University Bremen, dem Rochester Institute of Technology und der Universität Osnabrück tätig, von wo sie 2009 einen Masterabschluss in Informatik und 2006 einen Bachelorabschluss in Mathematik und Informatik erlangt hat. Ihre Forschungsschwerpunkte liegen im Bereich der Robotik, 3D-Umgebungskartierung, Laserscanningtechnologien, 3D-Punktwolkenverarbeitung und Multi-Sensor-Fusion. Ihre Forschungsarbeit ist wesentlicher Bestandteil der Open-Source Software “3DTK - The 3D Toolkit”, die 2009 mit dem Intevation Förderpreis für freie Software der Universität Osnabrück ausgezeichnet wurde.

# Anfragebeantwortung in Probabilistischen Datenbanken und Wissensbasen[1]

İsmail İlkan Ceylan[2]

**Abstract:** Probabilistische Datenbanken und Wissensbasen werden immer wichtiger in der Wissenschaft und Industrie. Sie werden ständig mit neuen Daten erweitert, angetrieben durch moderne Informationsextraktionssysteme, die Fakten mit Wahrscheinlichkeiten assoziieren. Der Stand der Technik, solche Daten zu speichern und zu verarbeiten, basiert auf probabilistischen Datenbanksystemen, die breit und erfolgreich eingesetzt werden. Jenseits von allen Erfolgsgeschichten fehlt solchen Systemen aber immer noch die grundlegende Maschinerie, um das in ihnen gespeicherte wertvolle Wissen an die Endnutzer weiterzugeben, was ihre potenziellen Anwendungen in der Praxis begrenzt. In ihrer klassischen Form basieren solche Systeme in der Regel auf starken, unrealistischen Einschränkungen, wie der *Welt-* und *Domänenabgeschlossenheit*, der *Tupelunabhängigkeit* und *dem Mangel an Allgemeinwissen*. Diese Einschränkungen führen nicht nur zu unerwünschten Konsequenzen, sondern setzen diese Systeme auch bei wichtigen Aufgaben, wie der Anfragebeantwortung, auf ein schwaches Fundament. In dieser Arbeit erweitern wir probabilistische Datenbanken und Wissensbasen mit realistischeren Datenmodellen und ermöglichen damit bessere Mittel für die Anfragebeantwortung. Aufbauend auf dem langen Bestreben, Logik und Wahrscheinlichkeit zu integrieren, entwickeln wir unterschiedliche Semantiken für probabilistische Datenbanken und Wissensbasen, analysieren ihre algorithmischen Eigenschaften und entwerfen, wann immer möglich, effiziente Anfragebeantwortungsalgorithmen.

## 1 Einführung

Es besteht ein starkes Interesse daran, große probabilistische Wissensbasen aus Daten auf automatisierte Weise aufzubauen, was zu einer Reihe von Systemen wie DeepDive [Sh15], NELL [Mi15], Reverb [FSE11], Microsoft Probase [Wu12], IBM Watson [Fe12] und Google Knowledge Vault [Do14] geführt hat. Diese Systeme durchsuchen kontinuierlich das Web und extrahieren *strukturierte* Informationen und füllen so ihre Datenbanken mit Millionen von Entitäten und Milliarden von Tupeln auf. Die Forschung auf dem Gebiet der großen Wissensbasen dient als neue Ära für die Integration von Logik und Wahrscheinlichkeit.

Inwieweit können diese Such- und Extraktionssysteme bei realen Anwendungen helfen? Obwohl sie sich noch in einem frühen Entwicklungsstadium befinden, werden Systeme wie DeepDive routinemäßig zum Aufbau von Wissensbasen für Bereiche wie *Paläontologie*, *Geologie*, *medizinische Genetik* und *menschliche Bewegung* eingesetzt; siehe, z.B., [Ku15] und [Pe14]. IBM Watson revolutioniert *Gesundheitssysteme* [Fe13] und viele andere Anwendungsgebiete der *Naturwissenschaften*. Google Knowledge Vault hat mehr als

[1] Originaltitel: Query Answering in Probabilistic Data and Knowledge Bases

[2] University of Oxford, ismail.ceylan@cs.ox.ac.uk (Nominierung bei der Technischen Universität Dresden).


(a) Mozart (b) Beethoven

Abb. 1: Informationsfelder für die Google-Suche (a) Mozart und (b) Beethoven

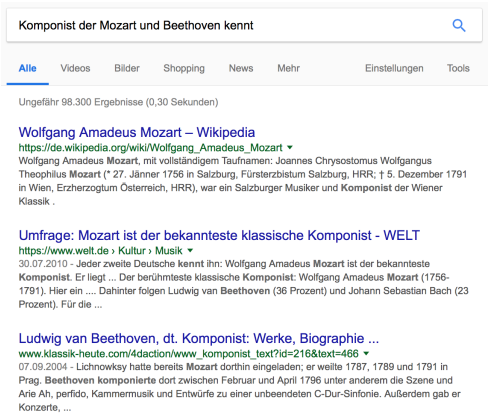
eine Milliarde Fakten aus dem Web zusammengestellt und wird hauptsächlich zur Verbesserung der Qualität von Suchergebnissen im Web verwendet.

Aus einem größeren Blickwinkel betrachtet, ist die Suche nach dem Aufbau großer Wissensbasen ein neuer Meilenstein für die Forschung im Bereich der künstlichen Intelligenz (KI). Bereiche wie Informationsextraktion, natürliche Sprachverarbeitung (z.B. Beantwortung von Fragen), relationales und tiefgehendes Lernen, Wissensrepräsentation und -schlussfolgerung und Datenbanken ergreifen Initiative für ein gemeinsames Ziel. Inferenzverfahren und die Anfragebeantwortung auf großen probabilistischen Wissensbasen wird allgemein als das Kernstück dieser Bemühungen angesehen. Vielleicht liegt die sichtbarste Anwendung von probabilistischen Wissensbasen in Suchmaschinen. Heutzutage wird die Standardliste relevanter Webseiten oft um eine Tabelle mit strukturierten Daten erweitert, die sich auf die Suchanfrage bezieht. Zum Beispiel zeigt die Suche nach Mozart (Figur 1a) oder Beethoven (Figur 1b) eine Box, die ihre Kinder, Ehepartner, Schüler, Geburtsorte usw. identifiziert, was eindeutig mit der zugrundeliegenden Wissensbasis verknüpft ist.

Neben allen Erfolgsgeschichten fehlt es den probabilistischen Wissensbasen jedoch immer noch an der grundlegenden Maschinerie, um einen Teil des wertvollen Wissens, das sich in ihnen versteckt, dem Endnutzer zu vermitteln [WHS16], was ihre potenziellen Anwendungen in der Praxis stark einschränkt. Zum Beispiel ist die Information, die neben den Suchergebnissen in der Google-Suche angezeigt wird, sehr stark moderiert und zeigt nur Fakten, über die der Suchmaschinenanbieter absolut sicher ist. Andere, unsicherere Informationen sind vor dem Benutzer verborgen.

Darüber hinaus gibt es keine Unterstützung für relationale Anfragen über diesen Datenbanken. Selbst die einfache Anfrage “Gibt es einen Komponisten, der sowohl Mozart als auch Beethoven kennt?” kann von diesen Systemen derzeit nicht beantwortet werden, trotz der Tatsache, dass es eine mögliche Antwort gibt: Haydn, eine Person, die der probabilis-


(a) Haydn




(b) Gibt es ein Komponist der sowohl Mozart als auch Beethoven kennt?

Abb. 2: Unfähigkeit, relationale Anfragebeantwortung auf großen Wissensbasen durchzuführen. Obwohl Mozart (Figur 1a), Beethoven (Figur 1b) und Haydn (Figur 2a) der probabilistischen Wissensbasis bekannt sind, kann die Anfrage, ob es einen Komponisten gibt, der sowohl Mozart als auch Beethoven kennt (Figur 2b), von diesem System nicht beantwortet werden.

tischen Wissensbasis bekannt ist (Figur 2a). Was macht die Auswertung einer solch einfachen Anfrage so schwierig, und warum kann dieses Wissen nicht an den Benutzer weitergegeben werden? Diese Fragen bilden unsere globale Motivation, und die Antworten sind mit tiefen theoretischen Problemen sowie mit technischen Einschränkungen verbunden, die wir als Nächstes skizzieren.

## 2 Probabilistische Datenbanken

Das grundlegendste Modell ist das der tupelunabhängigen probabilistischen Datenbanken (PDBs) [Su11], die tatsächlich vielen dieser Systeme zugrunde liegen. Wir betrachten ein (endliches) relationales Vokabular $\sigma$ bestehend aus *endlichen* Mengen **R** von *Prädikaten*, **C** von *Konstanten* und **V** von *Variablen*. Ein *Term* ist eine Konstante oder eine Variable; ein *Atom* hat die Form $\mathsf{P}(s_1,\ldots,s_n)$, wobei $\mathsf{P}$ ein $n$ -stelliges Prädikat ist, und $s_1,\ldots,s_n$ Terme sind. Ein *Grundatom* ist ein Atom ohne Variablen.

Eine Datenbank $\mathcal{D}$ ist eine endliche Menge von *Grundatomen*. Eine *probabilistische Datenbank* $\mathcal{P}$ ist eine endliche Menge von *(probabilistischen) Atomen* der Form $\langle t : p\rangle$, wobei $t$ ein Grundatom ist und $p \in [0,1]$, und immer wenn $\langle t : p\rangle, \langle t : q\rangle \in \mathcal{P}$, dann muss $(p = q)$ gelten. Eine PDB $\mathcal{P}$ ordnet jedem Atom $t$ die Wahrscheinlichkeit $p$ zu, wenn $\langle t : p\rangle \in \mathcal{P}$, und andernfalls die Wahrscheinlichkeit 0 zu. Unter der *Tupelunabhängigkeitsannahme* induziert eine solche Wahrscheinlichkeitszuordnung $\mathrm{P}$ die folgende *eindeutige gemeinsame Wahrscheinlichkeitsverteilung* über klassischen Datenbanken $\mathcal{D}$:

$$\mathrm{P}_{\mathcal{P}}(\mathcal{D}) := \prod_{t\in\mathcal{D}} \mathrm{P}_{\mathcal{P}}(t) \cdot \prod_{t\notin\mathcal{D}} (1-\mathrm{P}_{\mathcal{P}}(t)).$$

Aus Gründen der Effizienz fehlt probabilistischen Datenbanken typischerweise ein geeigneter Umgang mit Unvollständigkeit in der Praxis. Insbesondere kann jedes der oben genannten Systeme nur einen Teil der realen Welt modellieren, und diese Beschreibung ist zwangsläufig unvollständig. Wenn es jedoch um die Anfragebeantwortung geht, verwenden die meisten dieser Systeme starke und in den meisten Fällen unrealistische Vollständigkeitsannahmen, die ihre Anwendbarkeit einschränken. Wir werden nun einen Blick auf diesen Annahmen werfen, die inhärent mit der Semantik von probabilistischen Datenbanken verknüpft sind.

Aus modelltheoretischer Sicht basieren probabilistische Datenbanken auf Annahmen i) wie der *Weltabgeschlossenheit (WA)*, ii) der *Tupelunabhängigkeit (TU)*, iii) dem *Mangel an Allgemeinwissen (MA)* und iv) der *Domänenabgeschlossenheit (DA)*. Hier bedeutet WA, dass alle Fakten, die nicht in der probabilistischen Datenbank erscheinen, die Wahrscheinlichkeit 0 haben. Dies kann auch als eine probabilistische Variante der klassischen WA [Re78] gesehen werden. Die TU besagt, dass jedes Tupel in der probabilistischen Datenbank als unabhängige Bernoulli-Zufallsvariable interpretiert wird. MA bedeutet, dass es nicht möglich ist, explizites Domänenwissen zu kodieren. Und die DA impliziert, dass die betrachtete Domäne auf eine endliche Menge bekannter Konstanten festgelegt ist. Alle diese Annahmen kommen von klassischen Datenbanken, während die TU eine zusätzliche Annahme auf dem Wahrscheinlichkeitsraum ist. Wir konzentrieren uns zuerst auf die WA und illustrieren ihre Konsequenzen an einem einfachen Beispiel.

**Beispiel 1.** Wir betrachten eine tupelunabhängige PDB, die die Wahrscheinlichkeit 0.5 mehreren selbsterklärenden Fakten zuordnet:

$$\langle \mathsf{Komponist}(\mathsf{haydn}) : 0.5 \rangle, \langle \mathsf{LehrerVon}(\mathsf{haydn}, \mathsf{beethoven}) : 0.5 \rangle,$$
$$\langle \mathsf{Kennt}(\mathsf{haydn}, \mathsf{beethoven}) : 0.5 \rangle, \langle \mathsf{FreundVon}(\mathsf{haydn}, \mathsf{mozart}) : 0.5 \rangle.$$

Unter der GWA haben alle fehlenden Fakten die Wahrscheinlichkeit 0, dass heißt, sie sind falsch. Folglich erhalten die folgenden beiden Anfragen die Wahrscheinlichkeit **0**:

$$Q_1 = \exists x (\mathsf{LehrerVon}(x, \mathsf{beethoven}) \wedge \mathsf{GeborenIn}(x, \mathsf{österreich})),$$
$$Q_2 = \exists x (\mathsf{Person}(x) \wedge \neg \mathsf{Person}(x)).$$

Insbesondere wird angenommen, dass $\mathsf{GeborenIn}(\mathsf{haydn}, \mathsf{österreich})$ die Wahrscheinlichkeit 0 hat (d.h. falsch ist); jedoch kann diese Annahme selbst falsch sein. Tatsächlich kann $\mathsf{GeborenIn}(\mathsf{haydn}, \mathsf{österreich})$ sogar die Wahrscheinlichkeit 1 haben (d.h. kann wahr sein), was dazu führen würde, dass $Q_1$ die Wahrscheinlichkeit **0.5** hat.

Auf der anderen Seite ist $Q_2$ unerfüllbar und sollte immer die Wahrscheinlichkeit **0** haben, unabhängig davon, wie unvollständig die PDB ist. Das heißt, die GWA zwingt eine sehr flache Repräsentation, die es sogar unmöglich macht, eine erfüllbare Anfrage von einer unerfüllbaren zu unterscheiden. ◊

Wir können dieses Beispiel natürlich erweitern, um den Effekt der Tupelunabhängigkeitsannahme zu illustrieren.

**Beispiel 2.** Unter Tupelunabhängigkeit wird der Anfrage

$$Q_3 = \exists x\,(\mathsf{LehrerVon}(x, \mathsf{beethoven}) \wedge \mathsf{Kennt}(x, \mathsf{beethoven}))$$

die Wahrscheinlichkeit $0.5 \cdot 0.5 = \mathbf{0.25}$ zugeordnet. Aber da Haydn ein Lehrer von Beethoven ist, kennt er ihn auch; also sind die beiden Fakten nicht unabhängig. ◊

Diese Beobachtungen werden noch dramatischer, wenn mehrere Einschränkungen dieser Systeme kombiniert auftreten; insbesondere zusammen mit dem Mangel an Allgemeinwissen, der uns zu ontologischen Wissensbasen bringt.

## 3 Ontologisches Wissen und Probabilistische Wissensbasen

Der Mangel an Allgemeinwissen ist einer der Hauptgründe dafür, dass einige offensichtliche Antworten nicht aus den Wissensbasen abgerufen werden können. Dies zeigt sich in realen Anwendungen: Insbesondere im Kontext der Websuche, bei der die strukturierten Informationsergebnisse eindeutig mit der zugrundeliegenden Wissensbasis verknüpft sind.

**Beispiel 3.** Eine einfache Anfrage, ob es einen Komponisten gibt, der sowohl Mozart als auch Beethoven kennt,

$$Q_4 := \exists x\,\mathsf{Komponist}(x) \wedge \mathsf{Kennt}(x, \mathsf{beethoven}) \wedge \mathsf{Kennt}(x, \mathsf{mozart}),$$

erhält die Wahrscheinlichkeit 0 und kann daher von diesen Systemen nicht richtig ausgewertet werden. Die Antwort auf diese Frage ist tatsächlich in der Wissensbasis: Es ist bekannt, dass Haydn (i) ein Komponist, (ii) ein Freund von Mozart und (iii) einer der Lehrer von Beethoven ist.

In der Tat erhalten beide Anfragen "Freund von Mozart" und "Lehrer von Beethoven" die richtigen Informationen, die auf Haydn hindeuten, einen der Wissensbasis bekannten Komponisten. Allerdings fehlen explizite Informationen über $\mathsf{Kennt}(\mathsf{haydn}, \mathsf{mozart})$, und daher erhält diese Aussage die Wahrscheinlichkeit 0. ◊

Es ist schwierig, diese einfache Anfrage auszuwerten, weil die aktuellen PDBs kein Allgemeinwissen haben, nämlich, dass zwei befreundete Personen sich kennen, was ontologisch als

$$\forall x, y\,\mathsf{FreundVon}(x, y) \rightarrow \mathsf{Kennt}(x, y),$$

kodiert werden kann. Menschliches Schließen nutzt dieses grundlegende Wissen, um implizite Konsequenzen aus Daten abzuleiten, und diese Art von Wissen ist wesentlich für die Auswertung von Anfragen über großen PDBs in unkontrollierten Umgebungen wie dem Internet. Daher ist die Einbeziehung von Allgemeinwissen sehr wichtig, und dies ist inhärent damit verbunden, die obigen Vollständigkeitsannahmen von PDBs aufzugeben.

Die entscheidende Notwendigkeit, die Unabhängigkeitsannahme zu lockern, wurde bereits in mehreren neueren Ansätzen erkannt, die auf Markov-Logik-Netzen (MLNs) basieren [RD06, GS16]. All diesen Ansätzen ist gemeinsam, dass sie auch die Modellierung von logischem Wissen ermöglichen.

Anders als bei Ontologiesprachen verwenden MLNs jedoch die Annahme der Domänenabgeschlossenheit, was nicht immer vernünftig ist und zu problematischen (und sogar absurden) Konsequenzen über vielen nicht-trivialen Domänen führt. Wir veranschaulichen dies am folgenden Beispiel.

**Beispiel 4.** Berücksichtigen Sie das folgende ontologische Wissen:

$$\forall x\, \mathsf{Mensch}(x) \rightarrow \exists y\, \mathsf{Elternteil}(y,x)$$
$$\forall x,y\, \mathsf{Elternteil}(x,y) \rightarrow \mathsf{Vorfahr}(x,y)$$
$$\forall x,y,z\, \mathsf{Vorfahr}(x,y) \wedge \mathsf{Elternteil}(y,z) \rightarrow \mathsf{Vorfahr}(x,z)$$
$$\forall x\, \mathsf{Vorfahr}(x,x) \rightarrow \bot$$

Die erste Einschränkung besagt, dass *jeder einen Elternteil hat* und die anderen definieren die (azyklische) *Vorfahrenbeziehung*.

Wenn das Elternteil einer einzelnen Konstante $a$ in der Datenbank aufgrund der DA explizit nicht erwähnt wird, bedeutet diese Einschränkung in einem MLN, dass $a$ von den bekannten Konstanten ein Elternteil zugewiesen werden muss. Dies ist im Wesentlichen eine völlig zufällige Person in der Datenbank. Unter der Annahme, dass dieses Individuum ein Mensch ist, muss es auch einen Elternteil haben und so weiter. Da die Vorfahrenbeziehung jedoch azyklisch ist, bedeutet dies, dass mindestens eine bekannte Konstante einen Elternteil haben muss, das nicht menschlich ist. Während die Existenz eines solchen Individuums angesichts der Beschränkungen unvermeidlich ist, gibt es keine natürliche Grenze, die man auf die Anzahl der Menschen in der Datenbank setzen kann, bevor dies geschieht—die Domäne aller Menschen ist für alle praktischen Zwecke unbegrenzt.

Die DA, die in PDBs und MLNs verwendet wird, ist eindeutig nicht für unbegrenzte Domänen geeignet und kann nur Approximationen liefern. Zum Beispiel, selbst wenn eine feste Anzahl von Menschen zur Menge der Konstanten hinzugefügt wird, begrenzt dies effektiv die Anzahl der möglichen Generationen von Menschen (d.h. die Tiefe der Elternteil Relation) und die Anzahl der verschiedenen Menschen mit nicht-menschlichem Elternteil (die Anzahl der maximalen Elemente der Elternteil Relation). ◊

MLNs sind im Wesentlichen propositional und können keine unbekannten Individuen ausdrücken, was sie deutlich unterscheidet von vollwertigem prädikatenlogischem Wissen. Insgesamt ist die DA für viele Anwendungen, die von Natur aus über offene Domänen arbeiten, nicht geeignet.

Auf der anderen Seite sind Ontologien prädikatenlogische Theorien, die domänenspezifisches Wissen formalisieren, um dadurch automatisiertes Schließen zu ermöglichen. Ontologiesprachen operieren im Gegensatz zu MLNs über offenen Domänen. Die prominentesten Ontologiesprachen in der Literatur basieren auf Datalog$^{\pm}$ [CGP12, CGL12, CGK13]

und auf Beschreibungslogiken [Ba07]. Das Interpretieren von Datenbanken mit Allgemeinwissen in Form von Ontologien steht in engem Zusammenhang mit dem auf Ontologien basierenden Datenzugriff [Po08], der im Zusammenhang mit klassischen Datenbanken ausführlich untersucht wurde, um eine offene Welt- und Domänenanfragebeantwortung zu ermöglichen. In so einem Szenario wird eine Datenbankanfrage durch eine logische Schnittstelle vermittelt, um implizites Wissen explizit zu machen: das führt zu umfangreicheren Antworten für Anfragen.

## 4 Ein kurzer Blick auf die Ergebnisse

In dieser Dissertation [Ce17] erweitern wir probabilistische Wissensbasen mit realistischeren Datenmodellen und ermöglichen so bessere Antworten auf Anfragen. Wir entwickeln unterschiedliche Semantiken für probabilistische Datenbanken und Wissensbasen, analysieren ihre berechnungstechnischen Eigenschaften, und entwerfen, wann immer möglich, effiziente Abfragebeantwortungsalgorithmen. Um dies zu erreichen, bringt die aktuelle Arbeit einige neuere Paradigmen aus der Datenbanktheorie und Logik für die probabilistische Anfragebeantwortung und den damit verbundenen Inferenzaufgaben zusammen. Die Dissertation ist in vier Teile organisiert, wobei Teil I die Präliminarien einführt und Teil IV den Schlussfolgerungen gewidmet ist. Alle Ergebnisse sind in Teil II (Probabilistische Datenbanken) und Teil III (Logik und Probabilistische Wissensbasen) präsentiert, wie wir zunächst zusammenfassen.

### 4.1 Resultate für Probabilistische Datenbanken

Wir stellen probabilistische Datenbanken vor, definieren probabilistische Anfragebeantwortung als ein Entscheidungsproblem und untersuchen ihre berechnungstechnische Komplexität. Wir zeigen, dass die Datenkomplexitätsdichotomie des Berechnungsproblems (zwischen Polynomialzeit und #P [DS12]) auf das Entscheidungsproblem unter Turing-Reduktionen (zwischen Polynomialzeit und PP) übertragen werden kann. Über die bekannten Ergebnisse hinaus erhalten wir auch andere Komplexitätsergebnisse.

Anschließend stellen wir *offene probabilistische Datenbanken* vor, die als neues Datenmodell für probabilistische Datenbanken vorgeschlagen werden. Der Hauptunterschied zwischen den probabilistische Datenbanken und ihrer offenen Variante besteht darin, dass letztere die WA nicht anwenden. Wir bieten eine tiefe Diskussion über die semantischen Unterschiede zwischen diesen Modellen und vergleichen sie im Bezug auf die in dieser Arbeit identifizierten Ziele. Neben den semantischen Ergebnissen enthält dieses Kapitel auch eine gründliche Komplexitätsanalyse für eine Vielzahl von Anfragesprachen. Diese Analyse beinhaltet ein Dichotomie-Ergebnis für die Datenkomplexität (zwischen Polynomialzeit und PP) und einen effizienten Algorithmus (einen, der für die in Polynomialzeit berechenbaren Anfragen vollständig ist). Die Hauptergebnisse zu offenen probabilistischen Datenbanken wurden zuvor in [CDV16][3] veröffentlicht, und eine Kurzfassung dieser Arbeit erschien auch als eine eingeladene Publikation in [CDV17].

[3] Ausgezeichnet mit *Marco Cadoli Best Student Paper Prize* bei der KR 2016.

Wir untersuchen auch zwei alternative Inferenzprobleme für probabilistische Datenbanken; nämlich die Suche nach der *wahrscheinlichsten Datenbank* und der *wahrscheinlichsten Hypothese* für eine bestimmte Anfrage, die beide durch die *maximum a posteriori* Berechnungen von Probabilistischen Graphischen Modellen inspiriert sind. Wir argumentieren, dass diese Inferenzprobleme hilfreich sein können, um das volle Potenzial probabilistischer Datenbanken auszuschöpfen. Die meisten dieser Ergebnisse basieren auf der frühen Publikation [CBL17].

### 4.2 Resultate über Logik und Probabilistische Wissensbasen

Wir erweitern die Ergebnisse in Teil III, um auch Allgemeinwissen in Form von Ontologien einzubeziehen. Bei Ontologien werden Datenbankanfragen zusätzlich mit der Aussagekraft von Ontologien ausgestattet. Nach einer allgemeinen Namenskonvention in diesem Bereich bezeichnen wir solche Anfragen *ontologievermittelte Anfragen*. Wir untersuchen die probabilistische ontologievermittelte Anfragebeantwortung auf probabilistischen Datenbanken und ebenso auf offenen probabilistischen Datenbanken. Die Arbeit in diesen Abschnitten baut auf der früheren Veröffentlichung [BCL17] auf und bezieht sich auch auf [CLP16]. Schließlich betrachten wir die maximum a posteriori Probleme der probabilistische Datenbanken im Zusammenhang mit ontologievermittelten Anfragen, die auf früheren Arbeiten basieren [CBL17]. Alle diese Ergebnisse basieren auf Datalog$^{\pm}$-Ontologien, für die wir auch eine gründliche Komplexitätsanalyse anbieten.

Wir untersuchen auch die sogenannten Bayesschen Ontologiesprachen, die klassische Beschreibungslogiken mit probabilistischer Unsicherheit erweitern. Bayessche Ontologiesprachen wurden in [CP14a] vorgeschlagen und in [CP14b] weiter untersucht; anschliessend, kombiniert in einer Zeitschrift [CP17] als Teil eines Sonderheftes. Es besteht auch eine Proof-of-Concept-Implementierung für Schlussverfolgen in Bayesschen Ontologiesprachen [CMP15]. Unser Fokus ist auf ontologievermittelter Anfragebeantwortung (und den damit verbundenen Problemen), und wir bauen auf früheren Arbeiten auf [CP15b]. Darüber hinaus untersuchen wir auch einen neuartigen Monitoring-Ansatz, der die Macht der Ontologiesprachen mit dynamischen Bayesschen Netzen kombiniert; diese Kombination wurde zuerst in [CP15a] vorgeschlagen. Der resultierende Formalismus wird dann dynamische Bayesschen Ontologiesprachen genannt und erlaubt Projektionen über die zukünftigen Zustände eines Systems.

## Literaturverzeichnis

[Ba07] Baader, Franz; Calvanese, Diego; McGuinness, Deborah L; Nardi, Daniele; Patel-Schneider, Peter F, Hrsg. The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, 2nd. Auflage, 2007.

[BCL17] Borgwardt, Stefan; Ceylan, İsmail İlkan; Lukasiewicz, Thomas: Ontology-Mediated Queries for Probabilistic Databases. In: AAAI. 2017.

[CBL17] Ceylan, İsmail İlkan; Borgwardt, Stefan; Lukasiewicz, Thomas: Most Probable Explanations for Probabilistic Database Queries. In: IJCAI. 2017.

[CDV16] Ceylan, İsmail İlkan; Darwiche, Adnan; Van den Broeck, Guy: Open-World Probabilistic Databases. In: KR. 2016.

[CDV17] Ceylan, İsmail İlkan; Darwiche, Adnan; Van den Broeck, Guy: Open-World Probabilistic Databases: An Abridged Report. In: IJCAI. 2017.

[Ce17] Ceylan, İsmail İlkan: Query Answering in Probabilistic Data and Knowledge Bases. Dissertation, TU Dresden, 2017.

[CGK13] Calí, Andrea; Gottlob, Georg; Kifer, Michael: Taming the Infinite Chase: Query Answering under Expressive Relational Constraints. JAIR, 48:115–174, 2013.

[CGL12] Calí, Andrea; Gottlob, Georg; Lukasiewicz, Thomas: A General Datalog-Based Framework for Tractable Query Answering over Ontologies. JWS, 14:57–83, 2012.

[CGP12] Calí, Andrea; Gottlob, Georg; Pieris, Andreas: Towards More Expressive Ontology Languages: The Query Answering Problem. AIJ, 193:87–128, 2012.

[CLP16] Ceylan, İsmail İlkan; Lukasiewicz, Thomas; Peñaloza, Rafael: Complexity Results for Probabilistic Datalog$\pm$. In: ECAI. 2016.

[CMP15] Ceylan, İsmail İlkan; Mendez, Julian; Peñaloza, Rafael: The Bayesian Ontology Reasoner is BORN! In: ORE. 2015.

[CP14a] Ceylan, İsmail İlkan; Peñaloza, Rafael: The Bayesian Description Logic $\mathcal{BEL}$. In: IJCAR. 2014.

[CP14b] Ceylan, İsmail İlkan; Peñaloza, Rafael: Tight Complexity Bounds for Reasoning in the Description Logic BEL. In: JELIA. 2014.

[CP15a] Ceylan, İsmail İlkan; Peñaloza, Rafael: Dynamic Bayesian Ontology Languages. CoRR, abs/1506.08030, 2015.

[CP15b] Ceylan, İsmail İlkan; Peñaloza, Rafael: Probabilistic Query Answering in the Bayesian Description Logic BEL. In: SUM. 2015.

[CP17] Ceylan, İsmail İlkan; Peñaloza, Rafael: The Bayesian Ontology Language $\mathcal{BEL}$. JAR, 58(1):67–95, 2017.

[Do14] Dong, Xin; Gabrilovich, Evgeniy; Heitz, Geremy; Horn, Wilko; Lao, Ni; Murphy, Kevin; Strohmann, Thomas; Sun, Shaohua; Zhang, Wei: Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In: SIGKDD. ACM, 2014.

[DS12] Dalvi, Nilesh; Suciu, Dan: The dichotomy of probabilistic inference for unions of conjunctive queries. Journal of ACM, 59(6):1–87, 2012.

[Fe12] Ferrucci, D. A.: Introduction to "This is Watson". IBM J. Res. Dev. 56(3):235–249, 2012.

[Fe13] Ferrucci, David; Levas, Anthony; Bagchi, Sugato; Gondek, David; Mueller, Erik T.: Watson: Beyond jeopardy! AIJ, 199-200:93–105, 2013.

[FSE11] Fader, Anthony; Soderland, Stephen; Etzioni, Oren: Identifying Relations for Open Information Extraction. In: EMNLP. ACL, 2011.

[GS16] Gribkoff, Eric; Suciu, Dan: SlimShot: In-Database Probabilistic Inference for Knowledge Bases. Proceedings of VLDB Endowment, 9(7), 2016.

[Ku15] Ku, Joy P.; Hicks, Jennifer L.; Hastie, Trevor; Leskovec, Jure; Ré, Christopher; Delp, Scott L.: The mobilize center: An NIH big data to knowledge center to advance human movement research and improve mobility. Journal of the American Medical Informatics Association, 22(6):1120–1125, 2015.

[Mi15] Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; Krishnamurthy, J.; Lao, N.; Mazaitis, K.; Mohamed, T.; Nakashole, N.; Platanios, E.; Ritter, A.; Samadi, M.; Settles, B.; Wang, R.; Wijaya, D.; Gupta, A.; Chen, X.; Saparov, A.; Greaves, M.; Welling, J.: Never-Ending Learning. In: AAAI. 2015.

[Pe14] Peters, Shanan E.; Zhang, Ce; Livny, Miron; Ré, Christopher: A Machine Reading System for Assembling Synthetic Paleontological Databases. PLoS ONE, 9(12), 2014.

[Po08] Poggi, Antonella; Lembo, Domenico; Calvanese, Diego; De Giacomo, Giuseppe; Lenzerini, Maurizio; , Riccardo: Linking Data to Ontologies. JDS, 10, 2008.

[RD06] Richardson, Matthew; Domingos, Pedro: Markov Logic Networks. Machine Learning, 62(1):107–136, 2006.

[Re78] Reiter, Raymond: On closed world data bases. Logic and Data Bases, S. 55–76, 1978.

[Sh15] Shin, Jaeho; Wu, Sen; Wang, Feiran; De Sa, Christopher; Zhang, Ce; Ré, Christopher: Incremental Knowledge Base Construction Using DeepDive. Proceedings of VLDB Endowment, 8(11):1310–1321, 2015.

[Su11] Suciu, Dan; Olteanu, Dan; Ré, Christopher; Koch, Christoph: Probabilistic Databases, Jgg. 3. 2011.

[WHS16] Weikum, Gerhard; Hoffart, Johannes; Suchanek, Fabian: Ten Years of Knowledge Harvesting: Lessons and Challenges. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 39(3):41–50, 2016.

[Wu12] Wu, Wentao; Li, Hongsong; Wang, Haixun; Zhu, Kenny Q.: Probase: A Probabilistic Taxonomy for Text Understanding. In: SIGMOD. ACM, S. 481–492, 2012.

**İsmail İlkan Ceylan** hat Informatik an der Middle East Technical University in der Türkei studiert; danach wechselte er zum International Center for Computational Logic an der TU Dresden, wo er auch seinen Masterabschluss erhalten hat. Er promovierte unter der Leitung von Prof. Franz Baader am Lehrstuhl für Automatentheorie der TU Dresden. Während seiner Promotion absolvierte er einen dreimonatigen Forschungsaufenthalt im Automated Reasoning Lab, unter der Leitung von Prof. Adnan Darwiche an der University of California, Los Angeles. Außerdem hat er mehrere kürzere Forschungsaufenthalte an der University of Oxford gemacht, um mit Prof. Thomas Lukasiewicz zusammenzuarbeiten. Zurzeit arbeitet er als Forscher an der University of Oxford an dem EPSRC Projekt *"Realistic Data Models and Query Compilation for Large-Scale Probabilistic Databases"*, in dem er als Co-Investigator tätig ist.

# It's Getting Crowded! Verbesserung der Effektivität von Microtask Crowdsourcing

Ujwal Gadiraju[1]

**Abstract:** Microtask Crowdsourcing hat sich als gut geeignete Methode zur Erwerbung von menschlichem Input auf Abruf hervorgetan und findet verbreitete Anwendung für die Lösung zahlreicher Probleme. Bekannte Beispiele sind unter anderem Umfragen, das Erstellen von Inhalten und das Beschriften von Bildern. In den letzten zehn Jahren gab es bereits zahlreiche Microtask Crowdsourcing Anwendungen in verschiedenen Gebieten sowohl in der Forschung (von Sozialwissenschaften bishin zur Informatik) als auch für praktischen Nutzen in anderen Fachrichtungen. Dies hat ohne Frage die Grenzen von qualitativen und quantitativen Studien überschritten durch die Möglichkeit zuvor eingeschränkte Laborstudien und kontrollierte Experimente auszuweiten [HRZ11, PCI10]. Heutzutage lässt sich in kurzer Zeit und mit einfachen Mitteln ein Goldstandard für Evaluationen erstellen [GL10] und potentielle Teilnehmer mit unterschiedlichen Demographien sind rund um die Uhr erreichbar [Ip10, Di15]. Daraus ergeben sich jedoch auch eine Vielzahl von Herausforderungen, insbesondere im Hinblick auf die fehlende Kontrolle von Teilnehmern und die Qualität der erhobenen Daten. In dieser Arbeit beschäftigen wir uns mit einigen dieser Herausforderungen. Dabei liegt der Fokus jedoch nicht auf Anwendungen des ehrenamtlichen Crowdsourcing, wie etwa Citizen Science [Bo14], "Serious Games" oder "Games with a Purpose" [VAD08], Wikis [DRH11] oder ähnlichen Ansätzen. Stattdessen konzentrieren wird uns auf das Lösen einiger kritischer Probleme im Bereich des bezahlten Microtask Crowdsourcing, die überwunden werden müssen, um das volle Potential des Modells auszuschöpfen. Unsere Arbeit wird durch die Überzeugung beeinflusst und angetrieben, dass es immer Arbeitsströme geben wird, die von verschiedenen Teilen der Gesellschaft benötigt werden und nicht mittels ehrenamtlicher Teilnahme oder Gamification bearbeitet werden können und für die bezahlte Kanäle am besten geeignet sind, um die Anforderungen zu erfüllen [Ki13].

Einige der wesentlichen Herausforderungen im Bereich des Microtask Crowdsourcing, zu dem jeweiligen Zeitpunkt, an dem die verschiedenen Arbeiten in dieser Arbeit vorgestellt wurden, wurden durchgeführt und veröffentlicht wie im Folgenden beschrieben. Wir haben haben uns mit jeder dieser Herausforderungen methodisch auseinandergesetzt, wobei wir jeweils das aktuelle Verständnis von Crowd Work erweitert haben oder neue Lösungen vorgestellt haben, die die zu dieser Zeit existierenden Methoden leistungsmäßig übertroffen haben. Ziel dieser Arbeit ist die '*Verbesserung der Effektivität des Microtask Crowdsourcing Modells*. *Effektivität* ist in diesem Kontext definiert als Grad zu dem die Crowdworker hochqualitative Antworten beitragen und die Auftragsersteller die gewünschten Ergebnisse erhalten, wobei die Kosten (Aufgabenbearbeitungszeit, Bezahlung) für alle beteiligten Akteure optimiert werden sollen. Wir identifizieren in diesem Zusammenhang bislang offene Schlüsselherausforderungen und stellen Methoden vor, mit denen sich die durch die Probleme entstehenden Mängel überwinden lassen. Wir decken außerdem weitere Faktoren auf, die die von Crowdworkern produzierte Arbeit qualitativ beeinflussen und bislang nicht bekannt waren.

[1] L3S Research Center, Leibniz Universität Hannover, gadiraju@L3S.de

- **Herausforderung #1** *Eingeschränktes Verständnis von Crowdsourcing-Tasks und Worker Charakteristiken* — Mit einem Alter von etwa zehn Jahren ist das Microtask-Crowdsourcing Feld noch sehr jung. Zu verstehen, welche Aufgabentypen für Crowdsourcing infrage kommen, wird beim Entwickeln besserer Plattformen helfen und die Marketplace-Dynamiken zwischen Aufgabenerstellern und Crowdworkern verbessern. Als unsere Arbeit zu diesem Thema vor einigen Jahren durchgeführt wurde [Ga15], war wenig über das Verhalten von Crowdworkern auf Microtask Crowdsourcing Plattformen bekannt. Zu verstehen, wie sich Crowdworker in ihrem Verhalten unterscheiden und wie die Qualität der Arbeit von diesem Verhalten abhängt, kann den Aufgabenentwurf beeinflussen und zu effektiveren Mechanismen zur Qualitätskontrolle führen.

- **Herausforderung #2** *Mangelhafte Methoden zur Vorauswahl von Crowdworkern* — Oft existieren nur wenig oder gar keine Daten über die bisherige Leistung von Crowdworkern, die einen Hinweis auf die Qualität der Arbeit geben könnten. Um in Abwesenheit dieser Indikatoren Crowdworker mit den gewünschten Fähigkeiten und erwiesener Leistung auszuwählen, verwenden Aufgabensteller üblicherweise die Leistung von Crowdworkern in Qualifizierungstests oder in einem kleinen Teil der eigentlichen Aufgabe während der Pre-Screening Phase. Dies ist jedoch nur eine Annäherung an die Auswahl geeigneter Crowdworker. Benötigt werden deutlichere Indikatoren für die Kompetenz von Crowdworkern sowie effektivere Mechanismen für die Vorauswahl.
  Wie von Barry Schwartz in dessen einflussreichen Arbeiten im Bereich der Psychologie und Gesellschaftstheorie herausgestellt, führt ein Überangebot oft zu nachteiligen Effekten auf den Entscheidungsprozess von Menschen [Sc04, SW04]. Die große Auswahl an Aufgaben, die für einen erfahrenen Crowdworker auf einer großen Crowdsourcing Plattform (wie etwa Amazons Mechanical Turk (AMT)[3] oder Crowdflower[4]) zur Verfügung stehen, macht es schwierig, die Aufgaben zu finden, die am besten zum jeweiligen Crowdworker passen. Da viele Crowdworker sich für Aufgaben entscheiden, die nicht optimal zu ihnen passen, haben geeignetere Crowdworker aufgrund der Beschränkungen in der Anzahl der Teilnehmer nicht die Möglichkeit, an diesen Aufgaben zu arbeiten. Crowdworker nehmen oft an Aufgaben teil, die jenseits ihrer Fähigkeiten liegen, obwohl sie Interesse daran haben, ihren Ruf zu erhalten. Dadurch sinkt die Effektivität des Crowdsourcing-Ansatzes. Die Vorauswahl von Crowdworkern ist eine verbreitete Methode, um das Problem der Teilnahme ungeeigneter Crowdworker an einer Aufgabe zu lösen [Ol11].

- **Herausforderung #3** *Unvollständige Betrachtung von Faktoren, die qualitätsbezogene Ergebnisse beeinflussen* — Qualitätskontrolle ist vermutlich das meisterforschte Thema im Bereich des Microtask Crowdsourcing. Trotzdem werden die Aspekte, die die Qualität der produzierten Arbeit beeinflussen nicht vollständig betrachtet. Um eine faire und gerechtfertigte Behandlung der von auf einer Microtask Crowdsourcing Plattform erstellten Arbeit zu gewährleisten, ist es wichtig, alle qualitätsbeeinflussenden Faktoren zu verstehen und einzubeziehen. Dies

[3] `http://www.mturk.com/`
[4] `http://www.crowdflower.com/`

wird gegenwärtig wenig bis überhaupt nicht getan. Aufgabensteller und Plattformen berücksichtigen überlicherweise nur die von Crowdworkern produzierten Endergebnisse, ohne darauf zu achten, *wie* die Ergebnisse produziert wurden. So haben beispielsweise Arbeiten in der Etnographie, die sich mit dem Crowdsourcing-Prinzip auseinander gesetzt haben, deutliche Unterschiede in den Umgebungen, in denen sich verschiedene Crowdworker befinden, festgestellt [Gu14, Ma14].
Eine mangelnde Beschreibung der Aufgabe hat klare Konsequenzen: aus Mangel an Alternativen im Marketplace versuchen sich Crowdworker oft an Aufgaben, für die sie kein optimales Verständnis besitzen. Auf der anderen Seite sind Aufgabensteller sich oft nicht den Schwachstellen ihres Aufgabendesigns bewusst und halten unzufriedenstellende Ergebnisse als Beweise für bewusst schädliches Verhalten, so dass die Crowdworkern die Bezahlung für ihre Arbeit verweigern. Daraus resultiernd verlieren Crowdworker ihre Motivation, die Gesamtqualität der Arbeit sinkt und alle Akteure verlieren ihr Vertrauen in den Marketplace. Obwohl die Klarheit der Aufgabenstellung offensichtlich wichtig ist für das Microtask Crowdsourcing, gibt es kein klares Verständnis darüber, in welchem Umfang Unklarheit in Beschreibung und Anweisungen die Leistung der Crowdworker beeinflusst.

Wir haben wichtige Beiträge zu jeder der genannten Herausforderungen gebracht. Unsere Beiträge sind in Abbildung 1 dargestellt und weiter unten beschrieben.

- **Beitrag #1** *Erweiterung des aktuellen Verständnisses von Aufgabentypen, Crowdworker-Verhalten und Qualitätskontrolle* — Wir beschäftigen uns zunächst mit zwei zentralen Aspekten, die die Effektivität von Microtask Crowdsourcing beeinflussen: Aufgabendesign und das Verhalten von Crowdworkern. Um das aktuelle Verständnis von Microtasks und Crowdworker-Verhalten zu erweitern, haben wir eine weitfassende Studie mit 1000 Crowdworkern auf CrowdFlower durchgeführt.
  Basierend auf zuverlässigen Daten über die Aufgabentypen, die Crowdworker durchgeführt haben, haben wir ein zwei-Level Kategorisierungsschema für Microtasks vorgestellt und Einblicke in die Aufgabenbeliebtheit, den Aufwand von Crowdworkern für das Abschließen von verschiedenen Aufgabentypen und ihre Zufriedenheit mit der Bezahlung erhalten. Auf der obersten Ebene ist das Kategorisierungsschema *zielorientiert* (d.h. die Kategorien orientieren sich an den Aufgabenzielen) und die Sub-Kategorien sind *arbeitsflussorientiert* (d.h. die Sub-Kategorien orientieren sich an den Schritten, die durchgeführt werden müssen, um die Aufgabe erfolgreich abzuschließen). Die genaue Kategorisierung von Crowdsourcing-Aufgaben hat deutliche Auswirkungen auf die Nutzermodellierung von Crowdworkern und die Empfehlung von Aufgaben [GKD14]. Die vorgeschlagene Aufgabenkategorisierung, inklusive der Ergebnisse unserer Analyse, unterstützt Aufgabensteller bei Design und Durchführung der Aufgaben. Beispielsweise haben Difallah et al. in ihrer Analyse des AMT-Marketplace untersucht, wie verschiedene Typen von ”human intelligence tasks”(HITs) sich im Hinblick auf die von uns eingeführte Taxonomie über die Zeit weiterentwickelt haben [Di15]. Auf Basis unserer Ergebnisse im Zusammenhang mit aufgabenabhängigen Charakterisiken, wie etwa Aufgabenbeliebtheit, Aufgabenaufwand, benötigte Bezahlung usw. können Aufgaben-

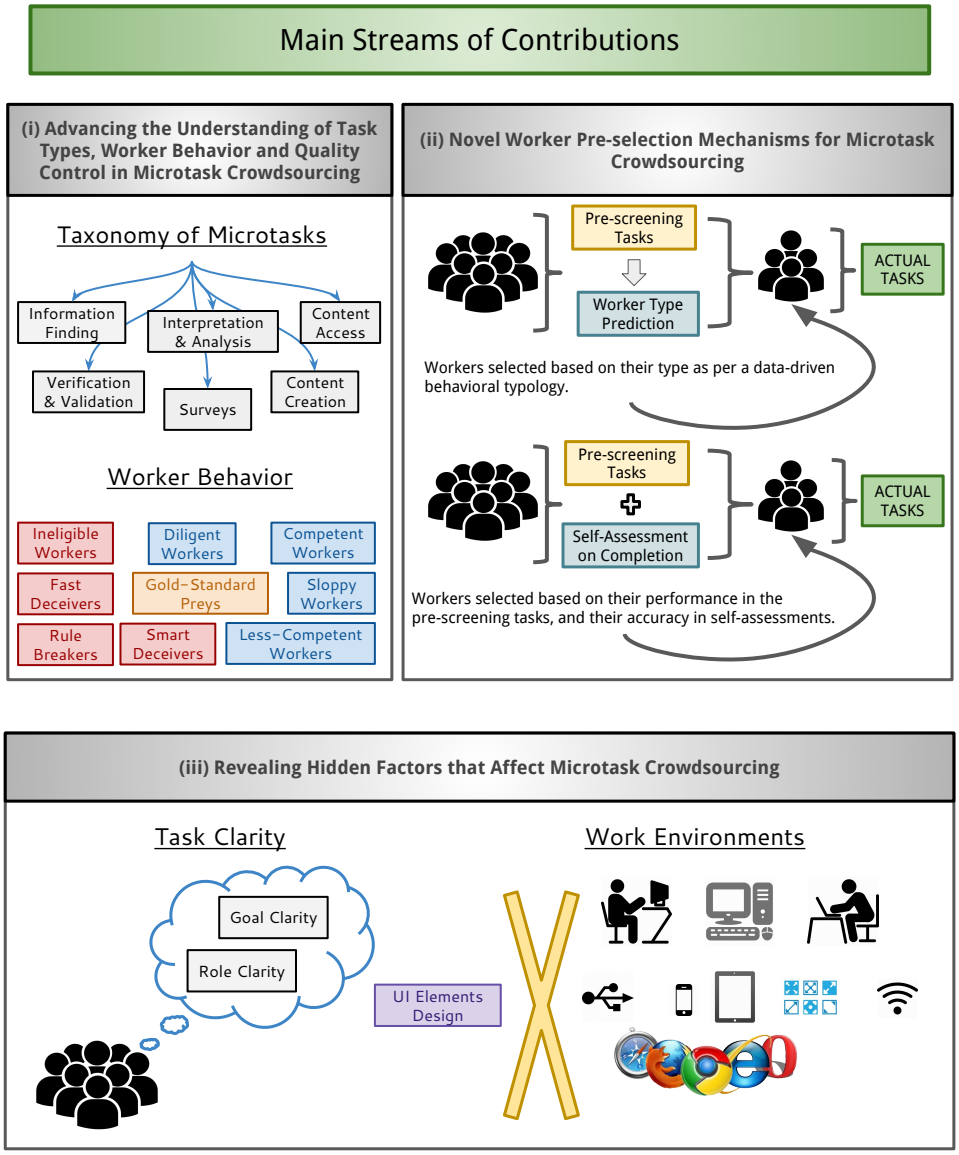



Abb. 1: Überblick über die wesentlichen Beiträge dieser Dissertation.

steller Aufgaben mit höherer erfolgsrate erstellen, d.h. die Qualität der Ergebnisse mit Rücksicht auf die Kosten maximieren.
Die Mechanismen zur Qualitätskontrolle müssen eine weite Gruppe unterschiedlicher Crowdworker und unterschiedlichen Verhaltens abdecken. Ein zentraler Schritt zum betrugsfreien Aufgabendesign ist das Verstehen der Verhaltensmuster von Microtask Crowdworkern. Wir haben die verbreiteten bewusst schädlichen Aktivitäten

auf Crowdsourcing-Plafformen analysiert und das gezeigte Verhalten von vertrauenswürdigen und nicht vertrauenswürdigen Crowdworkern studiert, insbesondere für Umfragen. Basierend auf unserer Analyse haben wir verschiedene Typen von schädlichem Verhalten identifiziert (*ineligible workers*, *fast deceivers*, *rule breakers* und *smart deceivers*), die über das hinaus gehen, was in bisherigen Arbeiten gezeigt wurde. Das Verstehen dieser Aspekte hilft uns, Aufgaben zu entwerfen, die schädlichem Verhalten entgegenwirken und somit den Aufgabenstellern sowie den Crowdsourcing-Plattformen nutzen.
Um die Qualität der Ergebnisse zu verbessern, haben wir Maße vorgestellt, die auf dem Verhalten der Crowdworker basieren und zum messen und entgegenwirken von ungewollten oder potentiell schädlichen Aktivitäten in Crowdsourcing-Aufgaben genutzt werden können. Wir haben eine detaillierte Analyse über schädliches Verhalten von Crowdworkern im Verlauf der Aufgabe vorgestellt und einen '*Wendepunkt*' definiert, der den Zeitpunkt markiert, ab dem ein Crowdworker dazu tendiert, schwache Antworten zu geben. In Anbetracht dieser Aspekte haben wir eine Anleitung zum effektiven Design von Crowdsourcing-Aufgaben eingeführt. Die Nutzung der Dynamik von Crowdsourcing-Aufgaben und das Prüfen von Crowdworker-Verhalten ist beim Design von besseren Aufgaben hilfreich.

- **Beitrag #2** *Neue Mechanismen für die Vorauswahl von Crowdworkern* — Wir stellen zwei verschiedene neue Methoden zur Vorauswahl von Crowdworkern vor, die die State-of-the-Art Ansätze für verschiedene Aufgabentypen leistungsmäßig übertreffen.
  Existierene Arbeiten haben die Aktivitäten von Crowdworkern anhand der jeweiligen Leistung binär in in *gut* und *schlecht* eingeteilt [RK11, DHL16]. Die Autoren dieser Arbeiten haben die Vorteile, die sich aus ihrem Ansatz verglichen mit anderen Qualitätskontrollmechanismen ergeben vorgestellt. Aspekte wie etwa Aufwand, Fähigkeiten und Verhalten können anhand der Aktivitäten der Crowdworker interpretiert werden und können bei der Einschätzung der Qualität der Arbeit des Crowdworkers helfen [RK11, RK12]. Obwohl es mit Sicherheit sinnvoll ist zwischen guter und schlechter Arbeitsqualität zu unterscheiden, sind wir der Meinung, dass eine feinere Unterscheidung der Aktivitäten von Crowdworkern weitere Vorteile mit sich bringt. Zum Beispiel führt das Wissen, dass selbst gute Crowdworker verschiedene Abläufe zur Bearbeitung von Aufgaben verwenden zu der Frage, ob diese Unterschiede praktische Auswirkungen haben können. Mit der zunehmenden Verwendung von Crowdsourcing-Lösungen in Form von menschlichem Input mittels Microtask-Marketplaces haben sich auch neue Anforderungen ergeben. Wenn verschiedene Beschränkungen der Kosten existieren (etwa bei Zeit oder Geld), genügt es oft nicht, die Qualität der Arbeit alleine vorherzusagen. Ein besseres Verständnis davon, wie gute Crowdworker sich in komplexen Aufgaben von anderen Crowdworkern abheben, kann zu Verbesserungen wie einem weiterentwickelten HIT-Design oder verbesserten HIT-Zuweisungsmodellen führen. Um das aktuelle Verständnis verschiedener Crowdworker-Typen auf einer Plattform zu vergrößern und dieses Verständnis für die Crowdworker-Vorauswahl bei gegebener Aufgabe zu nutzen, haben wir Daten über die Aktivitäten von Workern in 1800 HITs mit variierender Länge, Schwierigkeit und Typ erhoben. Wir

haben das Verständnis von Crowdworker-Typen neu definiert und es um multidimensionale Definitionen innerhalb der Typolody von Crowdworkern erweitert. Wir haben experimentell gezeigt, dass es möglich ist, Crowdworker automatisch in Klassen einzuteilen, basierend auf Machine Learning-Modellen, die die Verhaltensmuster von Crowdworkern bei der Arbeit an HITs verwenden. Die Nutzung einer Typ-Klassifizierung von Crowdworkern kann die Qualität von Crowdsourcing-Aufgaben durch die Vorauswahl von Workern für eine bestimmte Aufgabe verbessern. Unsere Vorauswahl-Methode führt zu einer Verbesserung von 10% verglichen mit bestehenden Vorauswahl-Verfahren. Wir stellen Verhaltensmerkmale für die Modellierung und Vorauswahl von Crowdworkern vor, basierend auf low-level Verhaltensmustern. Durch das Aufzeigen der praktischen Vorteile der Einteilung von guten Crowdworkern (*fleißige Worker* und *kompetente Worker*) nach unserer vorgestellten Typologie, haben wir gezeigt, dass eine Einteilung, die über *gut* und *schlecht* hinaus geht, wirksam ist. Dieses Ergebnis hat bedeutsame Auswirkungen auf Crowdsourcing-Systeme, in denen der Verhaltenstyp eines Crowdworkers vor der Teilnahme an einer Aufgabe nicht bekannt ist.
Auf der Basis von Selbsteinschätzungstheorien in der Psychologie, wie etwa dem Dunning-Kruger-Effekt, zeigen wir, dass Crowdworkern oft das Bewusstsein für ihr eigenes Kompetenzlevel fehlt und plädieren für kompetenzbasierte Vorauswahl in Crowdsourcing-Marketplaces. Der Dunning-Kruger-Effekt bezeichnet die konginitive Voreingenommenheit von weniger kompetenten Individuen, die zu einer überhöhten Selbsteinschätzung und illusionärer Überlegenheit führt [Du11]. Wir zeigen die Auswirkungen von mangelhafter Selbsteinschätzung auf reale Microtasks und stellen eine neue Vorauswahl-Methode für Crowdworker vor, die die Genauigkeit der Selbsteinschätzung von Crowdworkern miteinbezieht. Weiterhin haben wir verdeutlicht, dass die Einschätzung der Kompetenz eines Crowdworkers innerhalb einer Aufgabe durch die objektive Schwierigkeit der Aufgabe beeinflusst wird. Die Fähigkeit eine Crowdworkers, sich korrekt selbst zu beurteilen ist ein Teil der Kompetenz des Crowdworkers. Mittels gründlicher Bewertung für Tagging, Sentiment-Analysis und Image-Validation Aufgaben haben wir beobachtet, dass Aufgabensteller profitieren, indem sie die Selbsteinschätzung eines Crowdworkers mit in dessen Kompetenzbewertung miteinbeziehen, anstatt nur auf die Leistung des Workers in der Vorauswahl-Phase zu vertrauen. Unsere Ergebnisse zeigen, dass Worker, die mit unserem vorgestellten Ansatz ausgewählt werden, eine signifikant höhere Genauigkeit erreichen, verglichen mit Workern, die mittels herkömmlicher Vorauswahl-Methoden ausgewählt werden [Ga17b].

- **Beitrag #3** *Versteckte Crowdwork-Faktoren: Aufgabenklarheit und Arbeitsumgebung* — Worker auf Microtask-Crowdsourcing-Marketplaces streben nach einer Gleichgewicht zwischen finanziellem Einkommen und der Vergrößerung des eigenen Ansehen (Reputation). Dieses Gleichgewicht wird oft durch schlecht formulierte Aufgaben bedroht, da Crowdworker versuchen, diese durchzuführen, obwohl ihr Verständnis der Arbeit nicht ausreichend ist. Wir haben 100 Crowdworker auf der CrowdFlower-Plafform befragt, um das Vorhandensein von Problemen mit der Aufgabenklarheit in Crowdsourcing-Marketplaces zu verifizieren. Dabei haben wir herausgefunden, dass Crowdworker sich mit derartigen Problemen beschäftigen

müssen, was das Vorhandensein eines Mechanismus zur Vorhersage und Messung von Aufgabenklarheit motiviert. Inspiriert durch existierende Arbeit in Organisationspsychologie haben wir ein neues Modell für Aufgabenklarheit als Kombination von zwei Fragen vorgestellt: Wie klar ist das angestrebte Ziel einer Aufgabe (*goal clarity*)? Wie klar sind die durchzuführenden Aktivitäten (*role clarity*) [RN77]? Um besser zu verstehen, wie Klarheit von Crowdworkern wahrgenommen wird, haben wir die Einschätzungen von Crowdworkern für 7100 Aufgaben aus einem 5 Jahre umfassenden Datensatzes des AMT-Marketplaces gesammelt. In einer umfassenden Studie haben wir gezeigt, dass Klarheit von Workern zusammenhängend verstanden wird und vom Aufgabentyp abhängt. Zusätzlich haben wir Beweise dafür gefunden, dass Klarheit und Komplexität nicht direkt zusammenhängen. Dies zeigt eine komplexe Beziehung, die weiter untersucht werden sollte. Wir haben Features zum Erfassen von Aufgabenklarheit vorgestellt und die erhobenen Daten für das Trainieren und Validieren eines überwachten Machine Learning-Modells für die Vorhersage von Aufgabenklarheit verwendet. Anhand dieses Modells haben wir gezeigt, dass Aufgabenklarheit präzise vorhergesagt werden kann. Schließlich haben wir mittel einer temporalen Analyse gezeigt, dass Klarheit keine Macro-Eigenschaft eines AMT-Systems ist, sondern vielmehr eine lokale Eigenschaft, die durch Aufgaben und Aufgabensteller beeinflusst wird. Unsere Ergebnisse erweitern das aktuelle Verständnis von Crowdwork und haben wichtige Auswirkungen auf die Strukturierung von Arbeitsflüssen [GYB17]. Die Vorhersage von Aufgabenklarheit kann Crowdworkern bei der Aufgabenauswahl und Aufgabenstellern beim Design der Aufgaben helfen.

Ein weiterer Aspekt von Microtask-Crowdsourcing, der bisher wenig beachtet wurde, ist die *Arbeitsumgebung*, definiert als Hardware und Software, die von Crowdworkern für die Arbeit an Microtasks auf Crowdsourcing-Plattformen verwendet wird. Zunächst haben wir eine Pilotstudie zu den guten und schlechten Erfahrungen, die Crowdworker mit UI-Elementen in Crowdwork gemacht haben durchgeführt. Dabei haben wir die typischen Probleme aufgedeckt, mit denen sich Worker auseinandersetzen müssen. Wir haben herausgefunden, dass insbesondere folgende, schlecht designte UI-Elemente die Leistung von Crowdworkern negativ beeinflussen: große Eingabeboxen, unproportional kleine Textfelder und Multiple-Choice Fragen mit zu vielen Radio- bzw. Check-Boxen. Um den Einfluss verschiedener Designmöglichkeiten im Hinblick auf UI-Elemente, die Leistung von Crowdworkern und den Zusammenhang mit variierenden Worker-Umgebungen zu analysieren, haben wir eine zweite Studie mit über 125 verschiedenen Microtasks auf CrowdFlower durchgeführt, die sich in zwei identischen Teilen an Worker in Indien und den USA richtete. Diese Aufgaben haben die guten und schlechten Designs von UI-Elementen in Crowdsourcing-Aufgaben immitiert. Wir haben Hardwaredetails, wie CPU-Geschwindigkeit und Gerätetyp sowie Softwaredetails inklusive des verwendeten Browsers, Betriebssystems und weitere Eigenschaften, die die Arbeitsumgebung von Crowdworkern definieren festgehalten. Für Information-Finding und Content-Creation Aufgaben benötigen Crowdworker, die mobile Geräte benutzen, deutlich mehr Zeit als die Vervollständigung der Aufgabe im Vergleich zu Workern mit Laptops oder Desktop-PCs. Worker aus den USA waren im Durchschnitt

schneller und haben eine bessere Leistung bei Aufgaben mit schlecht designten UI-Elementen gezeigt, verglichen mit Workern aus Indien für alle Aufgabentypen und Aufgabenumgebungen. Außerdem haben Worker aus den USA indische Worker in Audi-Transcription Aufgaben leistungsmäßig übertroffen (bei einer gleichzeitig guten Leistung bei Aufgaben mit schlechter Audioqualität). Die Varianz an Arbeitsumgebungen ist bei US-Workern größer als bei indischen Workern und US-Worker verwenden öfter aktuelle Technologien (z.B. aktuelle Betriebssysteme und Browser). Um den Einfluss der Arbeitsumgebung auf Crowdsourcing-Microtasks besser zu verstehen, haben wir semi-strukturierte Interviews mit CrowdFlower-Workern, die alle Aufgaben bearbeitet haben, druchgeführt.
Durch unsere Studien haben wir die bedeutende Rolle von Arbeitsumgebungen auf Crowdwork offengelegt [Ga17a]. Unsere Ergebnisse deuten an, dass Crowdworker eine Vielzahl unterschiedlicher Arbeitsumgebungen verwenden, welche die Qualität der Arbeit beeinflussen. Wir haben herausgefunden, dass einige Arbeitsumgebungen Crowdworkers besser unterstützen als andere, in Abhängigkeit von den verwendeten UI-Elementen. Unser vorgestelltes Tool *ModOp* hilft beim Entwurf von Crowdsourcing-Microtasks, welche für unterschiedliche Arbeitsumgebungen geeignet sind. Wir haben empirisch gezeigt, dass der Einsatz von *ModOp* die kognitive Belastung von Workern reduziert und somit die Erfahrung der Worker verbessert. Die Nutzung reaktiver Nutzerschnittstellen, welche in der Lage sind, sich an verschiedene Arbeitsumgebungen anzupassen, würde vielen Crowdworkern die Möglichkeit geben, effektiver zu arbeiten. Unsere Ergebnisse haben außerdem wichtige Auswirkungen auf die Aufgabenzuteilung in Crowdwork. Aufgaben, die eine schnelle Durchführung voraussetzen oder von bestimmten Arbeitsumgebungen profitieren, können gezielt bestimmten Crowdworkern mit geeigneten Umgebungen zugewiesen werden.

Mit der Durchführung diverser interdisziplinärer Studien, der Vorstellung verschiedener Methoden zum Umgang mit den wesentlichen von uns identifizierten Herausforderungen und mit ausführlichen Evaluationen haben wir die folgenden nennenswerten Beiträge für die Verbesserung der Effektivität von Microtask-Crowdsourcing erbracht: (i) Wir haben das Verständnis von Aufgabentypen, Crowdworker-Verhalten und Qualitätskontrolle erweitert. (ii) Wir haben neue Mechanismen zur Vorauswahl von Crowdworkern vorgestellt, welche die existierenden Methoden leistungsmäßig übertreffen. (iii) Wir haben den Einfluss versteckter Faktoren, wie Aufgabeklarheit und Arbeitsumgebungen auf die Qualität der Arbeit herausgestellt.

## Literaturverzeichnis

[Bo14] Bonney, Rick; Shirk, Jennifer L; Phillips, Tina B; Wiggins, Andrea; Ballard, Heidi L; Miller-Rushing, Abraham J; Parrish, Julia K: Next steps for citizen science. Science, 343(6178):1436–1437, 2014.

[DHL16] Dang, Brandon; Hutson, Miles; Lease, Matt: MmmTurkey: A Crowdsourcing Framework for Deploying Tasks and Recording Worker Behavior on Amazon Mechanical Turk. arXiv preprint arXiv:1609.00945, 2016.

[Di15] Difallah, D. E.; Catasta, M.; Demartini, G.; Ipeirotis, P. G.; Cudré-Mauroux, P.: The Dynamics of Micro-Task Crowdsourcing – The Case of Amazon MTurk. In: 24th International Conf. on World Wide Web (WWW). 2015.

[DRH11] Doan, Anhai; Ramakrishnan, Raghu; Halevy, Alon Y: Crowdsourcing systems on the world-wide web. Communications of the ACM, 54(4):86–96, 2011.

[Du11] Dunning, David: The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance. Advances in experimental social psychology, 44:247, 2011.

[Ga15] Gadiraju, Ujwal; Kawase, Ricardo; Dietze, Stefan; Demartini, Gianluca: Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In: Proceedings of SIGCHI'15. S. 1631–1640, 2015.

[Ga17a] Gadiraju, Ujwal; Checco, Alessandro; Gupta, Neha; Demartini, Gianluca: Modus operandi of crowd workers: The invisible role of microtask work environments. Proceedings of the ACM on IMWUT, 1(3):49, 2017.

[Ga17b] Gadiraju, Ujwal; Fetahu, Besnik; Kawase, Ricardo; Siehndel, Patrick; Dietze, Stefan: Using Worker Self-Assessments for Competence-based Pre-Selection in Crowdsourcing Microtasks. ACM Transactions on Computer-Human Interaction (TOCHI), 24(4), 2017.

[GKD14] Gadiraju, Ujwal; Kawase, Ricardo; Dietze, Stefan: A taxonomy of microtasks on the web. In: Proceedings of the 25th ACM conference on Hypertext and social media. ACM, S. 218–223, 2014.

[GL10] Grady, Catherine; Lease, Matthew: Crowdsourcing document relevance assessment with mechanical turk. In: HLT-NAACL workshop on creating speech and language data with Amazon's mechanical turk. ACL, S. 172–179, 2010.

[Gu14] Gupta, Neha; Martin, David; Hanrahan, Benjamin V; O'Neill, Jacki: Turk-life in India. In: Proceedings of the 18th Int. Conf. on Supporting Group Work. S. 1–11, 2014.

[GYB17] Gadiraju, Ujwal; Yang, Jie; Bozzon, Alessandro: Clarity is a Worthwhile Quality – On the Role of Task Clarity in Microtask Crowdsourcing. In: Proceedings of the 28th ACM, HT '17, Prague, Czech Republic, July 4-7, 2017. ACM, 2017.

[HRZ11] Horton, John J; Rand, David G; Zeckhauser, Richard J: The online laboratory: Conducting experiments in a real labor market. Exp. Economics, 14(3):399–425, 2011.

[Ip10] Ipeirotis, Panagiotis G: Demographics of mechanical turk. 2010.

[Ki13] Kittur, Aniket; Nickerson, Jeffrey V; Bernstein, Michael; Gerber, Elizabeth; Shaw, Aaron; Zimmerman, John; Lease, Matt; Horton, John: The future of crowd work. In: Proceedings of CSCW'13. ACM, S. 1301–1318, 2013.

[Ma14] Martin, David; Hanrahan, Benjamin V; O'Neill, Jacki; Gupta, Neha: Being a turker. In: Proceedings of ACM CSCW'14. ACM, S. 224–235, 2014.

[Ol11] Oleson, David; Sorokin, Alexander; Laughlin, Greg P; Hester, Vaughn; Le, John; Biewald, Lukas: Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. Human computation, 11:11, 2011.

[PCI10] Paolacci, Gabriele; Chandler, Jesse; Ipeirotis, Panagiotis G: Running experiments on amazon mechanical turk. 2010.

[RK11] Rzeszotarski, Jeffrey M; Kittur, Aniket: Instrumenting the crowd: using implicit behavioral measures to predict task performance. In: Proceedings of the 24th annual ACM UIST. ACM, S. 13–22, 2011.

[RK12] Rzeszotarski, Jeffrey; Kittur, Aniket: CrowdScape: interactively visualizing user behavior and output. In: UIST'12. ACM, S. 55–62, 2012.

[RN77] Ruch, Libby O; Newton, Rae R: Sex characteristics, task clarity, and authority. Sex Roles, 3(5):479–494, 1977.

[Sc04] Schwartz, Barry: The paradox of choice: Why less is more. New York: Ecco, 2004.

[SW04] Schwartz, Barry; Ward, Andrew: Doing better but feeling worse: The paradox of choice. Positive psychology in practice, S. 86–104, 2004.

[VAD08] Von Ahn, Luis; Dabbish, Laura: Designing games with a purpose. Communications of the ACM, 51(8):58–67, 2008.

**Ujwal Gadiraju** hat seinen Bachelor-Abschluss in Informatik 2010 an der VIT University in Tamil Nadu, Indien erhalten. Seinen Master-Abschluss in Informatik hat er 2012 an der Delft Unversity of Technology in den Niederlanden erhalten. 2017 erhielt er seinen Doktor an der Fakultät für Elektrotechnik und Informatik an der Leibniz Universität Hannover. Zur Zeit arbeitet er als Postdoc am L3S Forschungszentrum an der Leibniz Universität Hannover. Seine wesentlichen Forschungsinteressen umfassen Human Computation, Crowdsourcing, Information Retrieval und Social Computing. Er erhielt den Douglas Engelbart Best Paper Award auf der ACM Conference on Hypertext and Social Media (HT) 2017, den Best Poster Award auf der ACM Web Science Conference (WebSci) 2016 und den Best Poster Award auf der International Semantic Web Conference (ISWC) 2014. Weiterhin wurde er 2017 mit dem Outstanding Reviewer Award auf der World Wide Web Conference (WWW) und der Excellent Reviewer Recognition auf der ACM CHI Conference on Human Factors in Computing Systems (SIGCHI) ausgezeichnet. Er hat über 40 wissenschaftliche Artikel veröffentlicht, darunter Papiere bei hochrangigen Konferenzen und einflussreichen Journalen wie ACM SIGCHI, ACM TOCHI, ACM HT, ACM UbiComp, WWW, ACM SIGIR, ACM CIKM, ISWC und anderen. In den vergangenen Jahren hat Ujwal aktiv zu europäischen Projekten beigetragen, wie etwa DURAARK und AFEL.

# Benutzerzentriertes Design für E-Partizipation

Katharina Große[1]

**Abstract:** Die hier präsentierte Forschung ermöglicht die benutzerzentrierte Entwicklung von E-Partizipationslösungen. Sie erarbeitet eine theoretisch fundierte Benutzertypologie, von der Anforderungen und Gestaltungsempfehlungen abgeleitet werden.

## 1 Relevanz und Forschungsziele

„Das Konstruieren von Systemen, das auf einem unangemessenen oder unvollständigen Verständnis der Erfordernisse der Benutzer beruht, ist eine der Hauptursachen für den Misserfolg von Systemen“ (DIN EN ISO 9241-210, S. 10). Der Einsatz von benutzerzentrierten Entwicklungsmethoden ist im privatwirtschaftlichen Bereich fast als Selbstverständlichkeit anzusehen. Im öffentlichen Sektor ist dies eine Seltenheit (siehe zu diesem Unterschied [Lo15]). Technologie hat ein nie dagewesenes Potential zur Stärkung der Demokratie und zur Verbesserung staatlicher Leistungen (siehe dazu beispielsweise [vL12]). Durch mangelnde Benutzerzentrierung wird dieses nicht vollständig ausgeschöpft. Dabei ist der Bedarf für verstärkte Benutzerzentrierung bei der Technologieentwicklung für den Staat sowohl Auftraggebern als auch Entwicklern bewusst. Bisher fehlen zur Umsetzung jedoch die notwendigen Ressourcen (siehe zum Bewusstsein und Ressourcenmangel [Gr18]).

Es gibt erste Bestreben, die Benutzerzentrierung in der IT-Entwicklung für den Staat durch Forschung zu unterstützen. Diese vernachlässigt bisher aber einen zentralen Aspekt: „To date, much of the HCI[2] research in the area has focused on the general user population, overlooking personality differences“ ([No13, S. 361]. Das ist problematisch, denn „häufig gibt es eine Anzahl unterschiedlicher Benutzergruppen . . . , deren Erfordernisse zu beachten sind“ (DIN EN ISO 9241-210, S. 14). Hier besteht eine besondere Herausforderung der Entwicklung für Politik und Verwaltung. Wie auf usability.gov [U.16] betont wird, können sich Leistungsangebote der öffentlichen Hand nicht auf bestimmte Benutzergruppen beschränken. Ihre Zielgruppe ist, kurz gesagt, jeder. Selbst wenn Ressourcen für die Recherche zu den Merkmalen und Erfordernissen der Zielgruppe vorhanden sind: Durch die Größe und Heterogenität der Benutzergruppe ist eine fundierte Erhebung kaum möglich.

Die hier vorgestellte Forschung legt die Grundlage, um benutzerzentrierte Entwicklung für Staat und Verwaltung dennoch zu ermöglichen. Am Beispiel von E-Partizipation identifiziert und beschreibt sie unterschiedlichen Benutzergruppen in einer Typologie. Darauf aufbauend leitet sie typspezifische Anforderungen und Gestaltungsempfehlungen für die Lösungsentwicklung ab.

[1] Zeppelin Universität, k.grosse@zeppelin-university.net
[2] Human-Computer Interaction, Mensch-Computer-Interaktion

## 2 Methodologie

Die hier vorgestellte Forschung ist explorativ. Die entwickelte Typologie stellt den ersten Schritt in einem benutzerzentrierten Gestaltungsprozess gemäß DIN EN ISO 9241-210 dar, dem „Verstehen und Beschreiben des Nutzungskontexts“ (S. 14). Darüber hinaus verfolgt die Arbeit einen gestalterischen Ansatz und leistet auch zum zweiten und dritten Schritt benutzerzentrierter Gestaltung einen Beitrag, zum „Spezifizieren der Nutzungsanforderungen“ und zum „Entwerfen der Gestaltungslösungen“ (S. 14). Diesem pragmatischen, gestalterischen Ansatz folgend, fließen Erkenntnisse aus quantitativer und qualitativer Forschung ein. In der erarbeiteten Merkmalsliste finden sich beobachtbare Konstrukte (beispielsweise Gerätebesitz oder Online-Verhalten) und solche, die nur durch Befragung oder Rekonstruktion zu erfassen sind (beispielsweise Motivationen, Eigenschaften, Einstellungen).

Die Typologie wurde in einem mehrstufigen Prozess erstellt (siehe Abbildung 1). Im ersten Schritt (1) wurde eine Liste von zu verwendenden Merkmalen ausgearbeitet. Dazu wurden die zentralen Theorien der politischen Partizipation und Technologienutzung analysiert. Für die politische Partizipation sind die Arbeiten von Milbrath und Goel [MG77] und Verba, Schlozman und Brady [VSB95] grundlegend. Auf die Frage nach Technologienutzung geben Venkatesh, Morris, Davis und Davis [Ve03] mit der United Theory of Acceptance and Use of Technology (UTAUT) eine umfassende Antwort. UTAUT wurde durch Venkatesh, Thong und Xu [VTX12] zu UTAUT2 weiterentwickelt und auf den Kontext der freiwilligen Nutzung angepasst. Sowohl bei Partizipation als auch bei Technologienutzung, wurden Betrachtungen zu psychologischen Aspekten in die Analyse aufgenommen. Auf der politischen Seite wurde zudem Forschung zum kollektivem Handeln einbezogen. Das Forschungsfeld befasst sich mit Einflüssen auf die individuelle Beitragsentscheidung bei gemeinsamen Gütern (beispielsweise [Kl04]). Auf der technischen Seite wurde die Forschung zu Nutzen und Belohnung hinzugenommen, die sich damit beschäftigt, wie und warum Menschen Medien nutzen (beispielsweise [KBG74]). Aus dieser Analyse ergab sich eine vorläufige Merkmalsliste.

Diese Merkmalsliste wurde validiert, indem die Merkmale mit den Variablen bestehender Studien zum Nutzungsverhalten bei Online-Partizipation abgeglichen wurden. Der Katalog der in die Validierung einbezogenen Studien wurde systematisch erstellt. Meckel, Hoffmann, Lutz, und Poell [Me14] bieten einen ausgezeichneten Einstiegspunkt, mit einer umfassenden Zusammenstellung aller deutsch- und englischsprachigen Artikel zum Thema Online-Partizipation in Journals mit Peer-Review-Verfahren, die im ISI Web of Knowledge, ProQuest, EBSCO oder der Mendeley-Datenbank verfügbar sind. Dies wurde mit einem Überblick von Mossberger [Mo09] ergänzt. Zusätzlich wurden alle relevanten Studien betrachtet, die in den ausgewählten Arbeiten referenziert werden.

Im nächsten Schritt (2) wurden Ausprägungscluster für die validierte Merkmalsliste identifiziert. Dazu wurden bereits existierende Typologien verwandter Forschungsfelder herangezogen, die Merkmalsüberschneidungen mit der entwickelten Liste aufweisen. Zwar gibt es eine Vielzahl von Typologien, die sich mit Online-Verhalten und teilweise auch Partizipationsaktivitäten beschäftigen. Diese beschränken sich jedoch meist auf Verhal-

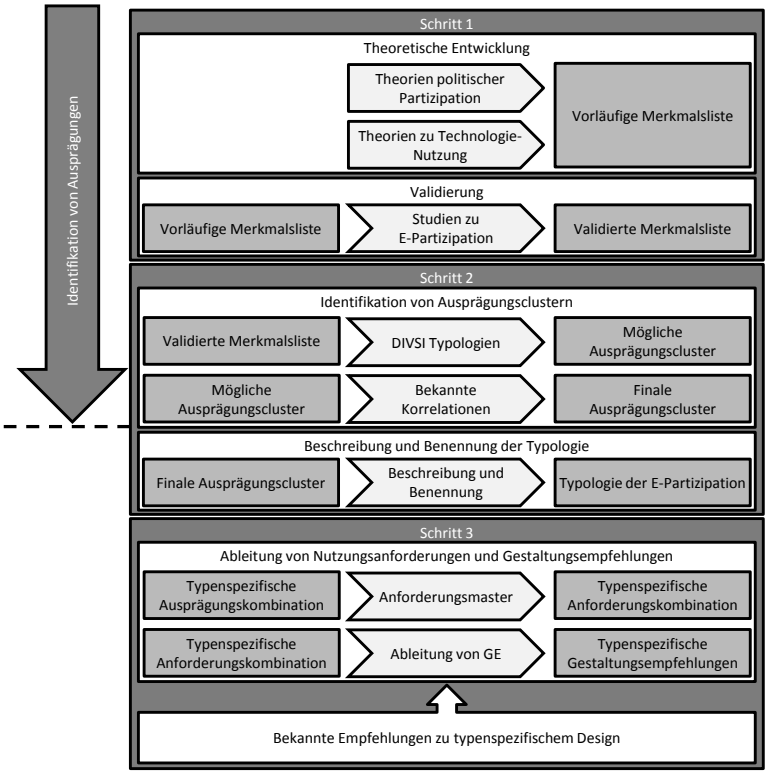



Abb. 1: Ablauf der Untersuchung. Quelle: Angepasst von [Gr18]

tenskonstrukte und teilweise auf sozio-demographische Variablen zur Beschreibung der Typen. Sie bieten kaum Überschneidungen mit der Merkmalsliste. Eine größere Schnittmenge weisen die DIVSI Internet Milieus auf. In diesen Milieus wird Internetverhalten und Einstellungen zum Internet verbunden mit dem „unterschiedlichen lebensweltlichen Hintergrund der Typen, d. h. Ihre Werthaltungen und Lebensstile“ [Si12, S. 10]. Diese Milieus wurden ergänzt durch eine Folgestudie, die sich detaillierter mit dem Internetverhalten der Milieus beschäftigt [HLP15] und durch eine gesonderte Betrachtung der unter 25-Jährigen [Si14].

Um diese drei Studien verwenden zu können, mussten die Konstrukte der DIVSI-Typologie der Merkmalsliste zugeordnet werden. Diese Übersetzung erforderte eine detaillierte Analyse der vorhandenen Informationen, angelehnt an das Verfahren der qualitativen Inhaltsanalyse [GL09]. Die danach vorhanden Lücken in den Clustern konnten durch die Analyse bekannter Ausprägungskorrelationen geschlossen werden.

Aus der so entstandenen Typologie der E-Partizipation konnten in Schritt 3 für jeden Typ Anforderungen und Gestaltungsempfehlungen abgeleitet werden. Letztere wurden mit Hilfe von Untersuchungen abgeleitet, die einzelne Design-Elemente mit Bezug auf Typologie-Merkmale analysieren.

## 3 Typologie der E-Partizipation

Es wurden fünf unterschiedliche Typen identifiziert: (1) Gestalter, (2) Optimierer, (3) Spieler, (4) Weltverbesserer, (5) Bemühte. Sie werden durch 44 Merkmalen beschrieben, die Angaben zum Online-Verhalten, zu Persönlichkeitsmerkmalen, Einstellungen zu Politik und IT sowie zu Motivationen der Typen enthalten. Für alle Typen wurde neben einer detaillierten tabellarischen Präsentation eine ausführliche Beschreibung anhand ihrer charakterisierenden Merkmale erstellt (siehe [Gr18]). Diese wird hier nur sehr verkürzt wiedergegeben. In diesem Beitrag stehen die abgeleiteten typspezifischen Design-Empfehlungen im Vordergrund.

Typ 1 (Gestalter) ist politisch interessiert und zeigt hohe Online-Fähigkeiten und ebenso hohe -Nutzungsintensität. Er möchte neue IT-Lösung ausprobieren und unterstützen. Gleichzeitig sieht er sich auch verpflichtet dazu, seinen Teil zur Gemeinschaft beizutragen. Typ 2 (Optimierer) wird politisch aktiv, wenn er bestimmte Ziele erreichen will. Das Internet wird zielgerichtet, bedarfsorientiert und sehr intensiv mit hoher Expertise genutzt. Das Verhältnis von Aufwand und Erfolg ist bei ihm zentral. Mobile Lösungen sind für ihn essentiell. Typ 3 (Spieler) sucht Zugehörigkeit und Erlebnisse. Er steht dem politischen System eher ablehnend gegenüber. Obwohl dieser Typ nur mittlere Online-Expertise aufweist, ist er im Internet zu Hause und begeistert sich für IT. Sein Aktionsradius hat einen stark hedonistischen Fokus. Typ 4 (Weltverbesserer) ist selbstbewusst und gewissenhaft, aber weniger offen für Neues. Er wird durch Altruismus und Pflichtgefühl motiviert. Seine Einstellung zu IT ist pragmatisch und er nutzt das Internet zweckorientiert und zielgerichtet mit mittlerer Intensität und Vielfalt bei mittlerer Expertise. Typ 5 (Bemühte) ist eher unsicher und introvertiert. Er schreibt sich selbst wenig Einfluss auf politische Geschehnisse zu und engagiert sich am ehesten aus einem Gefühl der gegenseitigen Verpflichtung heraus. Er setzt sich mit neuen Technologien nur auseinander, um nicht ausgegrenzt zu werden. Trotzdem freut er sich über Möglichkeiten, online seine Meinung äußern zu können.

## 4 Typspezifische Design-Empfehlungen

Insgesamt wurden 58 Anforderungen und 70 Gestaltungsempfehlungen abgeleitet. Für alle Typen ergab sich eine spezifische Kombination von Gestaltungsempfehlungen, die aus deren Anforderungs- und Merkmalskombinationen abgeleitet wurde. Die genaue Zusammensetzung der Merkmale, Anforderungen und Gestaltungsempfehlungen sind als detaillierte Grafiken und Beschreibungen in [Gr18] verfügbar. Die folgende Darstellung konzentriert sich auf die für die Typen charakteristischen Kombinationen.

### 4.1 Gestalter

Für die Gestalter ist das Internet ein Lebensraum. Sie sind es gewohnt, sich produzierend einzubringen, ihre Umgebung zu gestalten. Dabei verbinden sie Arbeit und Spaß. Sie

sind der einzige Typ, für den eine kollaborative Weiterentwicklung von Beteiligungsgegenständen (BG) ermöglicht werden sollte. Dies kann über eine Diskussionsseite gelöst werden, in die ein kollaborativer Text-Editor integriert ist. Ein zuschaltbares kollaboratives Whiteboard ermöglicht die Abweichung von textbasierter Interaktion und gestattet Gestaltern, externe Elemente einzubinden und zu bearbeiten. Der Gestalter ist ebenfalls der einzige Typ, für den die Gelegenheit geschaffen werden sollte, sich an der Weiterentwicklung des Systems zu beteiligen.

Zusätzlich wird für Gestalter eine offene Forumsstruktur empfohlen, sodass sie selbst BG einbringen können. So werden Benutzer zu Initiatoren und verwalten die Präsentationsseite, die es für jeden BG geben sollte. Sie wird ergänzt durch eine Wiki-Seite und sollte mit bestehenden redaktionellen Angeboten und offenen Daten verknüpfbar sein. Die Online-Beteiligung sollte außerdem mit einem parlamentarischen beziehungsweise einem Ratsinformationssystem (PIS/RIS) verbunden werden.

Für Gestalter ist es wichtig, dass auch „leisere Stimmen“ gehört werden. Deswegen bietet sich eine Funktion an, die wenig beachtete Beiträge anzeigt.Obwohl sie gerne neue Themen entdecken, möchten Gestalter zielgerichtet navigieren können, weshalb eine Suchfunktion und ein Themen-Speicher im Benutzer-Account empfehlenswert sind. Für Gestalter ist auch wichtig, dass Informationen verlässlich sind. Dies kann über gekennzeichnete Benutzer-Accounts für Auftraggeber gelöst werden. Die sichtbare Platzierung von Name und Logos des Auftraggebers trägt außerdem dazu bei, ein Gütesiegel zu verleihen. Zentral ist ein Fußnotensystem, bei dem Quellen für Informationen auf Präsentations- und Wiki-Seite angegeben werden können.

## 4.2 Optimierer

Optimierer sind effizienzbedacht und erfolgsorientiert. Sie sehen bei Online-Beteiligung ein schlechtes Kosten-Nutzen-Verhältnis. Ihnen muss ermöglicht werden, sich mit geringem Aufwand zu beteiligen. Ein mobiler Zugang ist für sie essentiell. Durch bisherige Erfolgsgeschichten oder die Anzeige der bisherigen Einflussrate von Vorschlägen auf den BG können sie vom Nutzen der Beteiligung überzeugt werden. Der Aufwand kann durch eine Auswahl zwischen Lösungsalternativen und eine Abstimmungsfunktion reduziert werden. Es wird außerdem empfohlen, in Infoboxen die wichtigsten Informationen prägnant zusammenfassen und auf der gleichen Seite eine Diskussionsfunktion einzurichten. Zum schnellen Diskussionseinstieg können Diskussionsverläufe visualisiert und zentrale Beiträge angezeigt werden. Zielgerichtete Navigation über selbstgepflegte Themen-Speicher und Suche sind zentral. Themen-Benachrichtigungen informieren Optimierer, wenn es für sie passende BG gibt. Eine Karten-Übersicht zeigt ihnen BG in ihrer Nähe. Automatisierte Text-Zusammenfassungen können angeboten werden, um bei langen Texten zentrale Aspekte wiederzugeben.

### 4.3 Spieler

Spieler sind kaum politisch interessiert und hauptsächlich hedonistisch motiviert. Die Online-Beteiligung muss als Erlebnis gestaltet sein, einen niederschwelligen Einstieg bieten und Spieler von der Wirksamkeit ihrer Beiträge überzeugen. Informationen sollten an einer zentralen Stelle präsentiert werden. Es wird empfohlen, für jeden BG in einem kurzen Video den Sinn und Einfluss der Online-Beteiligung zu erläutern. Ergänzt werden diese durch Infoboxen mit angeschlossener Diskussionsfunktion. Auch die Einbindung von externen Inhalten in diese Beiträge sollte ermöglicht werden. Mini-Games auf der Startseite und Beteiligung durch Alternativen-Auswahl sind ein geeigneter Einstieg. Badges an Beiträgen verdeutlichen deren Verwendung.

Daneben sollte Spielern ermöglicht werden, Gemeinschaften zu finden und aufzubauen. Das leisten Nachrichten und Chat-Funktionen und anpassbare Profile. Eine Event-Funktion unterstützt Spieler dabei, Beteiligungsformate eigenständig zu organisieren. Spielern muss ermöglicht werden, Beiträgen zu verfassen, die nicht auf die eigene Person zurückzuführen sind, weshalb anonyme Teilnahme und pseudonyme Benutzeraccounts empfohlen werden.

Spieler reagieren ablehnend auf Autorität, soziale Anreize und Selbstdarstellung. Es ist wichtig, die sich daraus ergebenden Widersprüche zu beachten. Erstens haben Spieler einerseits ein Bedürfnis, sich an Experten zu orientieren, lehnen aber die Autorität anerkannter Experten ab. Die Möglichkeit, beliebte Beiträge anzuzeigen löst diesen Konflikt. Sie weist Community-Experten aus, an deren Beiträgen sich Spieler orientieren können. Zweitens sollten Spieler auf der einen Seite für gute Beiträge belohnt werden. Auf der anderen Seite ruft die Betonung von erwünschtem Verhalten möglicherweise Trotz hervor. Dieser Widerspruch wird wie folgt gelöst: Funktionen, die Verhaltensanreize erzeugen sollen, aber keine spielerischen Elemente aufweisen, werden nicht empfohlen. Dazu zählen beispielsweise Nutzungsstatistiken. Freischaltbare Funktionen hingegen, appellieren primär an den Spieltrieb.

Spieler neigen dazu, sich im Internet treiben zu lassen. Sie werden nicht gezielt nach BG suchen. Es ist hingegen vorstellbar, dass sie auf einen Button klicken, der ihnen per Zufallsgenerator weitere Videos oder Benutzer-Kommentare anzeigt. Alles in allem muss in Bezug auf Spieler immer die Herausforderung gemeistert werden, auf ihre hedonistischen und sozialen Motivationen einzugehen, dabei aber trotzdem inhaltliche Beiträge zum BG zu fördern und möglicherweise politisches Interesse zu wecken.

### 4.4 Weltverbesserer

Für Weltverbesserer ist besonders wichtig, welche Wirkung von dem BG ausgeht, an dem sie sich beteiligen. Deshalb sollten die Relevanz und die Begünstigten des BG gut sichtbar präsentiert werden. Auf der gleichen Seite sollten den Weltverbesserern die Möglichkeit gegeben werden, über den BG zu diskutieren, ohne dass die Beiträge auf die eigene Person zurückgeführt werden können. Trotzdem möchten sie ihre Meinung darstellen, diese

mit ihrem Profil verknüpfen und andere Benutzer überzeugen. Es bieten sich pseudonyme Benutzeraccounts an. Für die Diskussion ist bei den Weltverbesserern wichtig, dass persönliche Konflikte vermieden werden. Eine Netiquette sollte erstellt und durchgesetzt werden.

Benutzer diesen Typs haben einige Bedenken, was Online-Partizipation angeht. Es sollte Vertrauen in die Sicherheit der Beteiligungslösung erzeugt und ein Gütesiegel für Informationen angeboten werden. Dazu bieten sich die Beschreibung des Sicherheitskonzepts an, sowie ein Fußnotensystem und gekennzeichnete Accounts für Akteure aus Politik und Verwaltung.

Für Weltverbesserer sollten Möglichkeiten geschaffen werden, Menschen zur Teilnahme an der Online-Partizipation einzuladen. Neue-Nutzer können per E-Mail zum Mitmachen aufgefordert werden. Für registrierte Benutzer sollte es eine Benutzer-Aufforderung zur Teilnahme an bestimmten BG geben.

Während die Leistung anderer Benutzer durchaus hervorgehoben werden sollte und Weltverbesserer auch Erfolg erleben möchten, sollte die Selbstdarstellung von Benutzer-Erfolgen auf ihren Profilen jedoch nicht betont werden. Weltverbesserer lehnen diese spezielle Form der Selbstdarstellung ab. Sie wollen sich zum Wohle Anderer beteiligen, nicht, um ihre Leistung in den Mittelpunkt zu rücken. Deswegen empfehlen sich zwar freischaltbare Funktionen und Nutzungsstatistiken, die auch ein gegenseitiges Pflichtgefühl auslösen. Erfolgsbadges hingegen, die beispielsweise die Qualität der eingereichten Vorschläge betonen, werden nicht empfohlen.

### 4.5 Bemühte

Bemühte möchten zwar ihre Meinung äußern, sich dabei aber nicht unbedingt selbst darstellen. Es geht ihnen eher darum, dass ihre Meinung überhaupt gehört wird. Bemühte haben niedrige Online-Fähigkeiten, nutzen nur wenige Online-Lösungen und sind skeptisch, was Veränderungen dieser Routinen betrifft. Bemühte sind aber auch, wie es ihr Name verdeutlichen soll, gewissenhaft, pflichtbewusst und altruistisch motiviert, weshalb sie trotzdem zu Online-Partizipation animiert werden können. Dies kann durch soziale Anreize geschehen, durch die Erzeugung eines Pflichtgefühls und die Beschreibung der Relevanz und der Begünstigten des BG. Auch die Einladung durch andere Teilnehmer ist für Bemühte ein Motivator. Sie müssen aber vom Mehrwert der Online-Partizipation überzeugt werden und den Einfluss der Online-Beteiligung auf das Endergebnis sowie die Verwendung der eigenen Beiträge nachvollziehen können. Der Wert der Benutzer-Beiträge muss betont werden und Bemühte sollten für gute Beiträge belohnt werden.

Sicherheitsbedenken müssen adressiert werden, unter anderem indem Beiträge nicht auf die Person zurückgeführt werden können. Bemühte sollten einfach an den gewünschten Punkt gelangen können und Informationen sollten an einer Stelle dargestellt werden. Die Beteiligungslösung sollte einfach zu bedienen und die Beteiligung ohne großen Aufwand möglich sein, während ein niederschwelliger Einstieg gewährleistet wird. Außerdem sollten sich Bemühte an Experten orientieren können.

Es wird somit eine offene Forumsstruktur empfohlen, bei der anonyme Teilnahme möglich ist. Es sind keine Benutzer-Accounts notwendig. Der Auftraggeber sollte beschrieben und klar durch Logos repräsentiert werden. Im Idealfall gibt es einen offiziellen Beteiligungsaufruf und Testimonials, die zur Beteiligung animieren. Das Sicherheitskonzept der Seite wird erläutert und Erfolgsgeschichten illustrieren den Einfluss, den Online-Beteiligung auf einen BG haben kann. Ein Investment-Zähler beschreibt, welche Leistung Teilnehmer bisher erbracht haben. Eine Karten-Übersicht erlaubt den Bemühten, BG zu identifizieren, die in ihrer Nähe angesiedelt sind. Die Anzeige von verwaisten Themen verdeutlicht den Bemühten, wo ihre Partizipation besonders gebraucht wird.

Der BG, seine Relevanz und Begünstigten, sowie der Mehrwert der Online-Beteiligung und der Wert und Einfluss der Beiträge werden in einem Video vorgestellt. Zusätzlich fassen Infoboxen wichtige Informationen zusammen. Unter dem Präsentationsvideo erlaubt eine einfache Kommentarfunktion den Bemühten, ihre Meinung darzustellen, ohne dass sie sich registrieren müssen. Die Durchsetzung einer Netiquette sorgt für einen angemessenen Ton in den Kommentaren. Durch die Alternativen-Auswahl und Bewertung von Kommentaren können sich Bemühte niederschwellig beteiligen. Sie können sich außerdem nur die beliebtesten Beiträge anzeigen lassen.

Durch die Präsentation von hochwertigen Kommentaren auf der Startseite und eine Feedback-Funktion können Bemühte belohnt werden und somit auch motiviert, inhaltliche Beiträge zu leisten. Eine Badge für die Beiträge kann verwendet werden, um zu kennzeichnen, wenn ein Kommentar Einfluss auf das Endprodukt genommen hat. Es wird empfohlen, dass Bemühten eine Funktion angeboten wird, die sie über E-Mail benachrichtigt, falls es ein Feedback oder einen Badge zu ihrem Kommentar gibt.

## 5 Diskussion der Ergebnisse

Die hier zusammengefassten Ergebnisse bieten eine wissenschaftlich fundierte Grundlage für die benutzerzentrierte Gestaltung von E-Partizipation. Im Idealfall würden sie über ein adaptives System umgesetzt, dass Benutzer in ein typologiebasiertes Benutzermodell einordnet. Solange dies nicht umgesetzt werden kann, muss in weiterer Forschung eine optimale Kombination von Gestaltungsempfehlungen für eine E-Partizipationslösung identifiziert werden. Wichtig ist dabei zu beachten, dass zwischen den Empfehlungen für einzelne Type durchaus Widersprüche auftreten können. Es lassen sich vier zentrale Konflikte identifizieren: (1) Funktionsvielfalt gegenüber Einfachheit, (2) Motivation durch soziale Anreize gegenüber Trotzreaktionen, (3) Vertrauenszuwachs durch offizielle Akteure gegenüber Ablehnung von Autoritäten, (4) Motivation durch inhaltliche Selbstdarstellung gegenüber Ablehnung dieser Selbstdarstellung.

Wichtig bei der Anwendung der Typologie ist vor allem, dass die erarbeiteten Anforderungen für die entsprechenden Typen erfüllt werden. Die Gestaltungsempfehlungen sind dafür hinreichend, aber nicht notwendig.Es ist außerdem sinnvoll, die hier vorgestellten Ergebnisse mit Ergebnissen aus der Forschung zu Unternehmensarchitektur-Modellen (Entreprise Architecture Frameworks) für E-Partizipation [SW12] und Design für unterschiedliche Nutzungsaufgaben (siehe für einen ersten Ansatz [PK08]) zu verbinden.

Die hier vorgestellte Arbeit zeigt, dass ein interdisziplinärer Ansatz zwischen Informationssystem- und sozial- beziehungsweise politikwissenschaftlicher Forschung große Mehrwerte für beide Disziplinen schafft. Sie schließt Forschungslücken im Bereich von HCI und Technologieakzeptanz für Politik und Verwaltung und verbessert das Verständnis von menschlichen Verhalten online. Gleichzeitig bietet sie durch den gestalterischen Ansatz einen wichtigen Erkenntnisgewinn für Entwickler von Lösungen für Online-Partizipation und trägt dazu bei, Nutzungs- und somit Beteiligungsbarrieren abzubauen.

## Literaturverzeichnis

[GL09] Gläser, Jochen; Laudel, Grit: Experteninterviews und qualitative Inhaltsanalyse. VS Verlag für Sozialwissenschaften, Wiesbaden, 3. Auflage, 2009.

[Gr18] Große, Katharina: Benutzerzentrierte E-Partizipation. Springer VS, Wiesbaden, 2018.

[HLP15] Hoffmann, Christian P.; Lutz, Christoph; Poell, Robin: , DIVSI Studie: Beteiligung im Internet - Wer beteiligt sich wie? https://www.divsi.de/wp-content/uploads/2015/07/DIVSI-Studie-Beteiligung-im-Internet-Wer-beteiligt-sich-wie_web.pdf, 2015.

[KBG74] Katz, Elihu; Blumler, Jay G.; Gurevitch, Michael: Uses and Gratification Research. The Public Opinion Quarterly, 37(4):509–523, 1974.

[Kl04] Klandermans, Bert: The Demand and Supply of Participation: Social-Psychological Correlates of Participation in Social Movements. In (Snow, David A.; Soule, Sarah A.; Kriesi, Hanspeter, Hrsg.): The Blackwell Companion to Social Movements, S. 360–379. Blackwell Publishing, Oxford, 2004.

[Lo15] Loutas, Nikolaos; Ojo, Adegboyega; Palmonari, Matteo; Paris, Cécile: Call for Papers: Special Issue on: Personalisation in e-Government and Smart Cities. International Journal of Electronic Governance, 2015.

[Me14] Meckel, Miriam; Hoffmann, Christian P.; Lutz, Christoph; Poell, Robin: , DIVSI-Studie zu Bereichen und Formen der Beteiligung im Internet. https://www.divsi.de/wp-content/uploads/2014/04/DIVSI-Studie-zu-Bereichen-und-Formen-der-Beteiligung-im-Internet.pdf, 2014.

[MG77] Milbrath, Lester W.; Goel, M. L.: Political Participation: How and Why Do People Get Involved in Politics? Rand Mc Nally College Publishing Company, Chicago, 2. Auflage, 1977.

[Mo09] Mossberger, Karen: Toward Digital Citizenship: Addressing Inequality in the Information Age. In (Chadwich, Andrew; Howard, Philip N., Hrsg.): Routledge Handbook of Internet Politics, S. 173–185. Routledge, New York, 2009.

[No13] Nov, Oded; Arazy, Ofer; López, Claudia; Brusilovsky, Peter: Exploring personality-targeted UI design in online social participation systems. Proceedings of CHI 2013, S. 361–370, 2013.

[PK08] Phang, Chee Wei; Kankanhalli, Atreyi: A Framework of ICT Exploitation for E-Participation Initiatives. Communications of the ACM, 51(12):128–132, 2008.

[Si12] Sinus Institut: , DIVSI Milieu-Studie zu Vertrauen und Sicherheit im Internet. https://www.divsi.de/sites/default/files/presse/docs/DIVSI-Milieu-Studie_Gesamtfassung.pdf, 2012.

[Si14] Sinus Institut: , DIVSI U25-Studie: Kinder, Jugendliche und junge Erwachsene in der digitalen Welt. https://www.divsi.de/wp-content/uploads/2014/02/DIVSI-U25-Studie.pdf, 2014.

[SW12] Scherer, Sabrina; Wimmer, Maria A: E-participation and enterprise architecture frameworks: An analysis. Information Polity, 17:147–161, 2012.

[U.16] U.S. Department of Health & Human Services: , Creating a User-Centered Approach in Government. http://www.usability.gov/what-and-why/user-centered-government.html, 2016.

[Ve03] Venkatesh, Viswanath; Morris, Michael G; Davis, Gordon B; Davis, Fred D: User acceptance of information technology: toward a unified view. MIS Quarterly, 27(3):425–478, 2003.

[vL10] von Lucke, Jörn: , Open Government - Öffnung von Staat und Verwaltung. http://www.zu.de/deutsch/lehrstuehle/ticc/JvL-100509-Open_Government-V2.pdf, 2010.

[vL12] von Lucke, Jörn: . Open Government Collaboration - Offene Formen der Zusammenarbeit beim Regieren und Verwalten. http://www.zu.de/deutsch/lehrstuehle/ticc/JvL-121025-OpenGovernmentCollaboration-V1.pdf, Friedrichshafen, 2012.

[VSB95] Verba, Sidney; Schlozman, Kay Lehmann; Brady, Henry E.: Voice and Equality: Civic Voluntarism in American Politics. Harvard Universtiy Press, Cambridge, MA, 1995.

[VTX12] Venkatesh, Viswanath; Thong, James Y. L.; Xu, Xin: Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. MIS Quarterly, 36(1):157–178, 2012.

**Katharina Große** forschte von 2012 bis Anfang 2017 am The Open Government Institute (TOGI) der Zeppelin Universität. Mit Abschluss ihrer Promotion wechselte sie in das Ministerium für Inneres, Digitalisierung und Migration Baden-Württemberg. Fokus ihrer Arbeit sind Benutzerfreundlichkeit und User Experience bei Electronic und Open Government.

# Software-basierte Mikroarchitekturangriffe[1]

Daniel Gruss[2]

**Abstract:** Moderne Prozessoren sind hoch optimierte Systeme, bei denen jeder einzelne Rechenzeitzyklus von Bedeutung ist. Viele Optimierungen hängen von den Daten ab, die verarbeitet werden. Mikroarchitekturangriffe greifen diese Daten unbefugt ab (Seitenkanäle) oder nutzen physikalische Unregelmäßigkeiten aus, um die Kontrolle über das gesamte System zu übernehmen (Fault-Angriffe). In meiner Dissertation [Gru17] habe ich den Stand der Technik bei Mikroarchitekturangriffen und Abwehrmechanismen in drei Dimensionen verbessert. Ich behandle diese kurz in dieser Zusammenfassung. Erstens zeige ich, dass Angriffe vollständig automatisiert werden können. Zweitens präsentiere ich einige neue, bisher unbekannte Seitenkanäle. Drittens zeige ich, dass Angriffe in stark eingeschränkten Umgebungen wie JavaScript in Websites sowie auf jedem Computersystem (Smartphones, Tablets, PCs und kommerzielle Cloud-Systeme) durchgeführt werden können.[3]

## 1 Einleitung

Die Idee, den Geheimcode für einen Safe zu lernen, indem man auf die Klickgeräusche des Schlosses hört, ist wahrscheinlich so alt wie Safes selbst. Das Klickgeräusch ist ein unbeabsichtigter Einfluss auf die Umgebung, der geheime Informationen preisgibt. 1996 beschrieb Kocher [Koc96] Seitenkanalangriffe, eine Technik, die es erlaubt, geheime Werte, die in einer Berechnung verwendet werden, aus unbeabsichtigten Einflüssen, die die Berechnung auf ihre Umgebung hat, abzuleiten. Diese bahnbrechende Arbeit war der Beginn eines ganzen Forschungsbereichs über Seitenkanäle. Kocher führte durch, was wir jetzt als Timing-Angriff bezeichnen – einen Angriff, der Unterschiede in der Ausführungszeit eines Algorithmus ausnutzt. In den folgenden Jahren wurden Seitenkanalangriffe basierend auf praktisch jeder messbaren Umgebungsänderung demonstriert, die durch verschiedene Arten von Berechnungen verursacht wurden. Beispiele sind etwa der Energieverbrauch, die elektromagnetische Abstrahlung, die Temperatur sowie photonische oder akustische Emissionen. All diese Angriffe haben jedoch gemeinsam, dass ein Angreifer physischen Zugriff auf das Zielgerät haben muss.

Im Gegensatz zu Seitenkanalangriffen, die dem Zielgerät keinen Schaden zufügen, gibt es auch Fault-Angriffe. Bei einem Fault-Angriff versucht ein Angreifer, die Berechnungen eines Geräts zu manipulieren, um Sicherheitsmechanismen des Geräts zu umgehen oder, um Geheimnisse des Geräts preiszugeben. Zu diesem Zweck manipuliert der Angreifer die Umgebung auf eine Weise, die das Zielgerät beeinflusst. Typischerweise können so Fehler induziert werden, wenn die Umgebung an den Rand des Spezifikationsbereichs

[1] Englischer Titel der Dissertation: “Software-based Microarchitectural Attacks”

[2] TU Graz, daniel.gruss@iaik.tugraz.at

[3] Die in meiner Dissertation veröffentlichten Ergebnisse spielten eine zentrale Rolle bei der Entdeckung der Meltdown [Li18] und Spectre [Ko19] Angriffe, die wir im Januar 2018 veröffentlicht haben.

des Zielgeräts oder darüber hinaus bewegt wird. Wie bei Seitenkanalangriffen wurden in der Vergangenheit verschiedene Umgebungsmanipulationen untersucht, z.B. Spannungsstörungen, Taktstörungen, extreme Temperaturen oder der Beschuss mit Photonen. Um einen Fault-Angriff auszuführen, ist ebenfalls eine Form von physischem Zugriff auf das Zielgerät erforderlich.

Moderne Computersysteme sind hochkomplex und hochoptimiert. Durch diese Optimierungen entstehen Informationslecks, unbeabsichtigte Einflüsse auf die Umgebung, die vom verarbeiteten Geheimnis abhängen. Optimierungen basieren auf spezifischen Datenwerten, die verarbeitet werden, dem Speicherort der Daten, der Häufigkeit von Zugriffen auf Speicherorte und vielen anderen Faktoren. Hier wird klar, dass ein Angreifer, der die Auswirkungen dieser Optimierungen durch einen Seitenkanal beobachtet, Schlüsse über die spezifische Ursache der Optimierungen ziehen kann. Das bedeutet, dass der Angreifer Informationen über die geheimen Datenwerte, die verarbeitet werden, lernt.

In meiner Dissertation [Gru17] habe ich softwarebasierte Mikroarchitekturangriffe untersucht. Softwarebasierte mikroarchitekturelle Seitenkanalangriffe nutzen Zeit- und Verhaltensunterschiede aus, die (teilweise) durch mikroarchitekturelle Optimierungen verursacht werden, d.h. Unterschiede, die nicht architekturell definiert oder dokumentiert sind. Softwarebasierte mikroarchitekturelle Fault-Angriffe nutzen physikalische Unregelmäßigkeiten im Gerät aus, um Fehler zu induzieren, d.h. sie operieren Elemente moderner Computersysteme am Rande oder außerhalb ihres Spezifikationsbereichs. Im Allgemeinen erfordern softwarebasierte Mikroarchitekturangriffe keinen physischen Zugriff, sondern nur eine Form der Codeausführung auf dem Zielsystem.

Zusammenfassend betrachten wir also Angriffe, die im Extremfall aus einer Website heraus sensitive Daten eines Systems abgreifen, ohne dass der Nutzer dies bemerkt (Seitenkanalangriffe), oder gar die komplette Kontrolle über das System übernehmen (Fault-Angriffe). Derartige Angriffe aufzuspüren ist essentiell, um die Angriffsfläche zu ermitteln und Gegenmaßnahmen zu entwerfen. Mikroarchitekturangriffe sind hoch-komplexe Angriffe, die weitreichende Grundlagen im Bereich der Betriebssystementwicklung und Prozessorarchitekturen benötigen. Während meiner Dissertation habe ich zu 13 Konferenz-Publikationen (davon sieben bei Top-Tier-Konferenzen) in diesem Themengebiet beigetragen von denen sechs (davon drei bei Top-Tier-Konferenzen) in meiner Dissertation vollumfänglich enthalten sind. Die Relevanz und den wissenschaftlichen Beitrag meiner Dissertation haben wir auch in der jüngeren Vergangenheit eindrucksvoll durch die Angriffe Meltdown [Li18] und Spectre [Ko19] bewiesen, für die meine Dissertation einen wichtigen Grundpfeiler darstellt.

## 2 Hintergrund

Cache-Angriffe sind die bekannteste Klasse softwarebasierter Mikroarchitekturangriffe. Die Möglichkeit, Timing-Unterschiede, welche durch Prozessor-Caches entstehen, für Angriffe auszunutzen, wurde erstmals von Kocher [Koc96] beschrieben. Cache-Timing-Angriffe wurden zuerst hauptsächlich auf kryptografische Algorithmen in softwarebasierten Angriffen angewendet.

Die meisten Cache-Angriffe in neueren Arbeiten sind Instanzen von drei generischen Cache-Angriffstechniken. Diese Techniken wurden bei gezielten Angriffen auf kryptographische Algorithmen verwendet und später von Osvik et al. [OST06] und Yarom et al. [YF14] verallgemeinert. Diese generischen Techniken sind unabhängig vom spezifischen Cache und der Hardware, auf der sie ausgeführt werden. Osvik et al. [OST06] beschrieben zwei generalisierte Cache-Angriffstechniken. Erstens den *Evict+Time*, bei der ein Angreifer misst, wie die Ausführungszeit eines Algorithmus beeinflusst wird, indem ein Cache-Set – eine Menge von kongruenten Cache-Zeilen – aus dem Cache entfernt wird. Zweitens, *Prime+Probe*, bei der ein Angreifer misst, ob eine Berechnung des Zielprogramms einen Einfluss darauf hat, wie lange es dauert, auf jeden Weg eines gewählten Cache-Sets zuzugreifen.

Bei beiden Angriffen erfährt der Angreifer, dass das gewählte Cache-Set vom Zielprogramm benutzt wurde. Yarom et al. [YF14] führten die dritte generische Angriffstechnik ein namens *Flush+Reload* ein. Bei einem *Flush+Reload*-Angriff löscht der Angreifer einen gemeinsam genutzten Speicherort aus dem Cache und misst anschließend, wie lange es dauert, um darauf zuzugreifen. Wenn das Zielprogramm in der Zwischenzeit den gemeinsam genutzten Speicher wieder in den Cache geladen hat, ist der erneute Zugriff schneller. Bei einem *Flush+Reload*-Angriff erkennt der Angreifer nicht nur, welches Cache-Set vom Zielprogramm verwendet wurde, sondern auch den exakten Speicherort (mit einer Genauigkeit der Cache-Zeilen-Größe).

Basierend auf diesen drei Angriffsprimitiven wurden verschiedenste Berechnungen angegriffen, zum Beispiel kryptografische Algorithmen [YF14], Webserver Funktionsaufrufe [ZJRR14] und Betriebssystemoperationen [HWH13]. Es war jedoch unklar, ob mit diesen Angriffen auch sensitive Informationen wie Nutzerpasswörter und Nutzereingaben ausspioniert werden können.

Alle Cache-Angriffe, die vor meiner Dissertation publiziert wurden, erforderten einen signifikanten Anteil an manueller Arbeit eines Experten, um den Angriff zu entwickeln und durchzuführen. Dies steht im Gegensatz zu klassischer Malware, die voll automatisiert agiert. Eine interessante Fragestellung ist also, inwieweit sich Cache-Angriffe automatisieren lassen. Auf modernen Prozessoren funktionieren außerdem die vor meiner Dissertation publizierten *Prime+Probe*-Angriffe nicht mehr. Die Herausforderung hier lag also darin, Angriffe so weit zu automatisieren, dass sie sich auch auf modernere Systeme automatisch anpassen können.

Softwarebasierte Fault-Angriffe sind in der Praxis erheblich schwieriger durchzuführen, da hier Fehler in der Hardware induziert werden müssen. Daher muss die Software eine Systemkomponente so weit an den Rand ihres Spezifikationsbereichs bringen – oder darüber hinaus –, sodass ein Fehler auftritt. Erst im Jahr 2014 haben sich softwarebasierte Fault-Angriffe als praktisch erwiesen, in dem sogenannten Rowhammer-Angriff [Ki2014; SD15]. Folglich war eine offene Frage die Anwendbarkeit von Rowhammer-Angriffen, d.h. ob Rowhammer beispielweise aus einer Website heraus durchgeführt werden könnte.

# 3 Beiträge der Dissertation zum Stand der Wissenschaft

Um mögliche Gegenmaßnahmen gegen softwarebasierte Mikroarchitekturangriffe zu entwickeln und zu bewerten, ist es notwendig, die Angriffsfläche detailliert abzubilden und zu verstehen. In meiner Dissertation [Gru17] habe ich dazu beigetragen, das allgemeine Verständnis der Angriffsfläche softwarebasierter Mikroarchitekturangriffe zu verbessern und habe neue Einblicke in softwarebasierte Mikroarchitekturangriffe und Angriffsvektoren geliefert. Meine Forschung umfasst die Minimierung von Anforderungen, die Automatisierung früherer Angriffe und die Identifizierung bisher unbekannter Seitenkanäle.

Ich begann mit der Arbeit an softwarebasierten Mikroarchitekturangriffen, indem ich die *Flush+Reload* Cache-Angriffstechnik [YF14] verbesserte. Frühere Cache-Angriffe erforderten die manuelle Identifikation von Schwachstellen (z.B. spezifische Datenzugriffe oder die Ausführung von Anweisungen, abhängig von geheimen Informationen). In meiner Dissertation [Gru17] habe ich die Technik **Cache-Template-Angriffe** vorgestellt. Cache-Template-Angriffe ermöglichen es, cache-basierte Informationslecks jedes Programms ohne vorherige Kenntnis bestimmter Softwareversionen oder Systeminformationen automatisch zu identifizieren und auszunutzen. Cache-Template-Angriffe können online auf einem Remote-System ohne vorherige Offline-Berechnungen oder Messungen ausgeführt werden. Ich habe diese Technik auf der USENIX Security 2015 Konferenz [GSM15] in Zusammenarbeit mit Raphael Spreitzer und Stefan Mangard veröffentlicht. Cache-Template-Angriffe bestehen aus zwei Phasen. In der Vorbereitungsphase bestimmen wir Abhängigkeiten zwischen der Verarbeitung von geheimen Informationen – z.B. bestimmten Schlüsseleingaben oder privaten Schlüsseln von kryptografischen Primitiven – und spezifischen Cache-Zugriffen. In der Angriffsphase leiten wir die geheimen Werte basierend auf beobachteten Cache-Zugriffen ab.

Wir haben verschiedene Anwendungen von Cache-Template-Angriffen untersucht. Unser automatisierter Angriff auf die T-Tabellen-basierte AES-Implementierung von OpenSSL ist genauso effizient wie die manuellen Cache-Angriffe nach dem Stand der Technik. Unsere Ergebnisse zeigen jedoch auch, dass ein Angreifer sowohl unter Linux als auch unter Windows, die Zeitpunkte sowie weitere Informationen aller Tastenanschläge mit hoher Genauigkeit mitprotokollieren kann. Für Linux-Distributionen demonstrierten wir sogar einen vollautomatischen Keylogger, der die Entropie von Passwörtern von $\log_2(26) = 4{,}7$ Bits pro Zeichen auf 1,4 Bits pro Zeichen signifikant reduziert. Daraus können wir schließen, dass cache-basierte Seitenkanalangriffe eine noch größere Bedrohung für die heutigen Computerarchitekturen darstellen als bisher angenommen. Tatsächlich können selbst sensible Benutzereingaben wie Passwörter, auf Maschinen die Prozessor-Caches verwenden, nicht als sicher betrachtet werden. Grundlegende Konzepte von Computerarchitekturen und Betriebssystemen ermöglichen diese automatische Ausnutzung von cache-basierten Sicherheitslücken. Unsere Beobachtungen haben gezeigt, dass viele der bestehenden Gegenmaßnahmen solche Angriffe nicht wie erwartet verhindern. Insbesondere reicht es nicht aus, nur bestimmte kryptographische Algorithmen wie AES zu schützen. Um der Bedrohung durch automatisierte Cache-Angriffe entgegenzuwirken, müssen wir generische Gegenmaßnahmen finden. Die Tatsache, dass Cache-Angriffe automatisch gestartet werden

können, stellt einen Perspektivwechsel dar, von einem eher akademischen Interesse hin zu praktischen Angriffen, die von weniger versierten Angreifern gestartet werden können.

Während Caches den vergleichsweise langsamen Arbeitsspeicher puffern, puffert der Arbeitsspeicher selbst die noch langsamere Festplatte. Daher sind Seitenkanalangriffe auch auf der Arbeitsspeicher-Ebene möglich. Suzaki et al. [SIYA11] demonstrierte einen Seitenkanal-Angriff auf Seitendeduplizierung, die vom Betriebssystem oder Hypervisor durchgeführt wird. Mit diesem Angriff kann festgestellt werden, ob ein Zielprogramm bestimmte Daten im Speicher hält. Folglich wurde Seitendeduplizierung in öffentlichen Clouds als schädlich betrachtet, aber in einer privaten Umgebung, z.B. privaten Clouds, PCs und Smartphones, immer noch als sicher eingestuft. Wir haben als Erste gezeigt, dass **Seitendeduplizierungsangriffe** sogar **in JavaScript** ausgeführt werden können. Im Gegensatz zu früheren Angriffen erfordert unser Angriff nicht, dass der Zielrechner das Programm des Angreifers ausführt, sondern einfach, dass eine Webseite auf dem Zielrechner geöffnet wird, die den JavaScript-Code des Angreifers enthält. Mit diesem Angriff lässt sich nicht nur feststellen, welche Anwendungen laufen, sondern es lassen sich auch bestimmte Benutzeraktivitäten beobachten, zum Beispiel ob der Benutzer bestimmte Webseiten geöffnet hat. Der Angriff funktioniert auf Servern, PCs und Smartphones und über die Grenzen von virtuellen Maschinen hinweg. Unsere Ergebnisse wurden auf der ESORICS 2015 Konferenz [GBM15] in Zusammenarbeit mit David Bidner und Stefan Mangard veröffentlicht. Die Ergebnisse unserer Arbeit zeigen, dass Systeme mit aktivierter Seitendeduplizierung nicht mehr allgemein als sicher betrachtet werden können. Die Tatsache, dass Seitendeduplizierungsangriffe über Websites gestartet werden können, stellt außerdem einen Paradigmenwechsel dar: von einem gezielten Angriff auf ein bestimmtes System hin zu groß angelegten praktischen Angriffen, die gleichzeitig auf einer großen Anzahl von Geräten ausgeführt werden können.

Basierend auf den beiden zuvor beschrieben Arbeiten untersuchte ich Möglichkeiten **Rowhammer-Angriffe** [Ki2014; SD15] **in JavaScript** durchzuführen. Rowhammer steht im Widerspruch zur grundlegenden Sicherheitsannahme, dass ein Speicherort nur durch Prozesse verändert werden kann, die darauf schreiben dürfen und dies auch tun. Parasitäre Effekte im Arbeitsspeicher können jedoch den Inhalt einer Speicherzelle ändern, ohne darauf zuzugreifen, indem auf andere Speicherstellen mit einer hohen Frequenz zugegriffen wird. Dieser sogenannte Rowhammer-Fehler tritt in den meisten heutigen Speichermodulen auf und hat fatale Folgen für die Sicherheit aller betroffenen Systeme, z.B. für Angriffe mit denen ein einfaches unprivilegiertes Programm Root-Privilegien erhält. Alle vorherigen Rowhammer-Studien und Angriffe beruhten auf der Verfügbarkeit der Cache-Flush-Anweisung, um Zugriffe auf Arbeitsspeicher-Module mit einer ausreichend hohen Frequenz zu verursachen.

Wir überwinden diese Einschränkung, indem wir komplexe Cache-Ersetzungsalgorithmen besiegen. Unsere Ergebnisse zeigen, dass ein Angreifer mit regulären Speicherzugriffen Caches sehr effizient dazu bringen kann, Daten aus dem Cache zu entfernen, und zwar effizient genug, um dann den Rowhammer-Fehler auszulösen. Unsere Arbeit ist die erste Arbeit, die Cache-Ersetzungsalgorithmen aus der Angreiferperspektive genauer untersucht und infolge die komplexen Cache-Ersetzungsrichtlinien erfolgreich umgeht. Dies ermöglicht nicht nur die Implementation von Rowhammer in JavaScript, sondern auch eine bessere

Erforschung von Cache-Angriffen, da nun Angriffe auf aktuelle und unbekannte CPUs schnell und zuverlässig ausgeführt werden können. Früher publizierte Gegenmaßnahmen schützen gegen diesen neuen Rowhammer-Angriff nicht.

Unser vollautomatischer Angriff läuft in JavaScript über eine Website und kann uneingeschränkten Zugriff auf Systeme erhalten. Die Angriffstechnik ist unabhängig von der CPU-Mikroarchitektur, der Programmiersprache und der Ausführungsumgebung. Da reguläre Computersysteme mit DDR3-Modulen und DDR4-Modulen anfällig sind, ist es wichtig, alle Angriffsvektoren von Rowhammer zu finden. Automatisierte Angriffe durch Websites stellen eine enorme Bedrohung dar, da sie auf Millionen von Zielmaschinen gleichzeitig ausgeführt werden können. Die Ergebnisse dieser Forschung wurden auf der DIMVA 2016 Konferenz [Gr16b] in Zusammenarbeit mit Clémentine Maurice und Stefan Mangard veröffentlicht.

Vorgeschlagene Gegenmaßnahmen gegen Cache-Angriffe treffen die Annahme, dass Cache-Angriffe mehr Cache-Hits und Cache-Misses verursachen als gutartige Anwendungen. Daher werden Hardware-Performance-Counter zur Erkennung verwendet. Um zu zeigen, dass diese Annahme nicht gilt, habe ich einen neuen Cache-Angriff namens ***Flush+Flush*** entwickelt. Der *Flush+Flush*-Angriff nutzt lediglich die Ausführungszeit der Flush-Anweisung aus, die davon abhängt, ob Daten im Cache sind. Der *Flush+Flush*-Angriff führt im Gegensatz zu anderen Cache-Angriffen keine Speicherzugriffe aus. Somit verursacht der Angriff überhaupt keine Cache-Misses und die Anzahl der Cache-Hits wird aufgrund der konstanten Cache-Flushes auf ein Minimum reduziert. Aus demselben Grund löst *Flush+Flush* keine Prefetches aus und ist daher in mehr Situationen als andere Angriffe anwendbar. Da der Angriff keine Cache-Fehler verursacht, schlagen Erkennungsmechanismen, die auf Hardware-Performance-Counter zur Überwachung der Cache-Aktivität setzen, fehl, da ihre zugrunde liegende Annahme falsch ist. Der *Flush+Flush*-Angriff kann in einer höheren Frequenz ausgeführt werden und ist daher schneller als jeder vorherige Cache-Angriff. Mit 496 KB/s in einem verdeckten Kanal, der CPU-Kern-übergreifend funktioniert, ist er 6,7 mal schneller als jeder zuvor veröffentlichte verdeckte cache-basierte Kanal. Um den *Flush+Flush*-Angriff zu verhindern, haben wir kleine Hardware-Modifikationen vorgeschlagen: Hätte die `clflush`-Anweisung eine konstante Laufzeit, würden keine messbaren Auswirkungen auf heutige Software entstehen, gleichzeitig der *Flush+Flush*-Angriff aber verhindert werden. Daher ist es eine wirksame Gegenmaßnahme, die implementiert werden sollte. Die Ergebnisse unserer Arbeit wurden auf der DIMVA 2016 Konferenz in Zusammenarbeit mit Clémentine Maurice, Klaus Wagner und Stefan Mangard veröffentlicht. Die Experimente in diesem Papier erweiterten das Verständnis der Interna moderner CPU-Caches. Neben den Ergebnissen zu Erkennungsmechanismen profitiert das Feld der Cache-Angriffe von Erkenntnissen rund um die `clflush`-Anweisung.

Intel x86-CPUs haben in der wissenschaftlichen Gemeinschaft eine beachtliche Aufmerksamkeit erhalten und es wurden leistungsfähige Techniken zum Ausnutzen cache-basierter Kanäle entwickelt. Moderne Smartphones verwenden jedoch eine oder mehrere Multicore-ARM-CPUs, die eine andere Cache-Organisation und einen anderen Befehlssatz als Intel x86-CPUs haben. Diese ARM-CPUs haben normalerweise keine Flush-Anweisung, die für reguläre Programme zugreifbar ist und im Gegensatz zu Intel x86-CPUs teilen sie

auch die Last-Level Caches nicht. Folglich wurden noch keine CPU-Kern-übergreifenden Cache-Angriffe auf nicht gerooteten Android-Smartphones gezeigt. Um hier den Stand der Forschung voranzutreiben, haben wir die wichtigsten Herausforderungen, die diese Angriffe bislang verhindert haben, gelöst und ***Prime+Probe*, *Flush+Reload*, *Evict+Reload* und *Flush+Flush* auf nicht gerooteten ARM-basierten Geräten** ohne jegliche Privilegien demonstriert. Unsere Angriffe sind die ersten kernübergreifenden und prozessorübergreifenden Angriffe auf ARM-CPUs. Basierend auf unseren Techniken haben wir verdeckte Kanäle gebaut, die vorherige verdeckte Kanäle auf Android um mehrere Größenordnungen übertreffen. Darüber hinaus bieten unsere Angriffstechniken eine hohe Auflösung und eine hohe Genauigkeit, die es ermöglicht, einzelne Ereignisse wie Touch- und Swipe-Aktionen auf dem Bildschirm, Touch-Aktionen auf der Soft-Tastatur und die Zeitabstände zwischen Tastenanschlägen zu überwachen. Folglich können wir auch die Länge der auf dem Touchscreen eingegebenen Wörter ableiten. Schließlich zeigen wir den ersten Cache Angriff auf eine kryptografische Primitive, die in Java implementiert ist. Wir zeigen, dass effiziente Angriffe gegen die Standard-AES-Implementierung, die Teil des Java-Bouncy-Castle-Crypto-Providers ist, durchgeführt werden können. Wir zeigen auch, dass die Cache-Aktivität der ARM TrustZone von der normalen Welt aus überwacht werden kann. Unsere Ergebnisse wurden auf der USENIX Security 2016 Konferenz [Li16] in Zusammenarbeit mit Moritz Lipp, Raphael Spreitzer, Clémentine Maurice und Stefan Mangard veröffentlicht. Wir sind davon überzeugt, dass mit unserem neuen Angriff auf Bibliotheken und Apps zahlreiche weitere ausnutzbare Informationslecks aufgedeckt werden. Unsere Angriffe sind auf Hunderte von Millionen von heute verfügbaren Standard-Smartphones anwendbar, da sie alle sehr ähnliche, wenn nicht sogar identische Hardware haben. Dies ist besonders beängstigend, da Smartphones zu den wichtigsten persönlichen Computern geworden sind und unsere Techniken den Umfang und die Auswirkungen von Cache-Angriffen erheblich erweitern.

Um neue unbekannte Seitenkanäle zu finden, habe ich das Verhalten von Prefetch-Anweisungen untersucht. Hierbei habe ich festgestellt, dass diese Anweisungen verwendet werden können, um moderne Betriebssysteme anzugreifen. Moderne Betriebssysteme verwenden Hardware-Unterstützung, um sich vor Kontrollfluss-Manipulation und Code-Manipulation zu schützen. Beispielsweise werden Schreibzugriffe auf ausführbare Seiten verhindert, und die Ausführung im Kernelmodus ist nur auf Kernel-Code-Seiten beschränkt. Gegenwärtige CPUs bieten jedoch keinen Schutz gegen Code-Recycling-Angriffe wie ROP (return-oriented programming). ASLR (Address-Space-Layout Randomization) wird verwendet, um diese Angriffe zu verhindern, indem alle Adressen für einen Angreifer unvorhersehbar gemacht werden. Somit basiert die Kernelsicherheit unter anderem auch auf dem Verhindern des Zugriffs auf Adressinformationen.

Im Rahmen meiner Dissertation [Gru17] entwickelte ich **Prefetch-Seitenkanalangriffe**, eine neue Klasse von generischen Angriffen, die gravierende Schwächen in Prefetch-Anweisungen ausnutzen. Die Timing-Unterschiede bei Prefetch-Anweisungen stammen von einer zweiten Cache-Hierarchie in modernen Prozessoren für Seitentabelleneinträge, die neben der regulären Cache-Hierarchie existiert. Ich habe herausgefunden, dass die Ausführungszeit von Prefetch-Anweisungen vom Zustand dieses Caches zur Seitenübersetzung abhängt. Ein noch gravierenderes Problem ist, dass die x86-Prefetch-Anweisungen

unprivilegierten Prozessen erlauben, privilegierten Speicher in den Cache zu laden. Damit erlauben diese neuen Angriffe nicht-privilegierten lokalen Angreifern, die Zugriffskontrolle auf Adressinformationen vollständig zu umgehen und somit ein gesamtes physisches System zu kompromittieren, indem sie Sicherheitsmechanismen wie SMAP, SMEP und Kernel-ASLR umgehen. Prefetch-Seitenkanalangriffe funktionieren in nativen und virtualisierten Umgebungen gleichermaßen.

Wir haben zwei Primitive eingeführt, die die Basis unserer Angriffe bilden. Erstens das Übersetzungsebenenorakel, welches die Ausführungszeit von Prefetch-Anweisungen ausnutzt. Zweitens, das Adressübersetzungsorakel, welches die fehlenden Privilegien-Überprüfungen ausnutzt.

Das Übersetzungsebenenorakel ermöglicht ASLR zu besiegen und Bibliotheken und Treiber in unzugänglichen Speicherbereichen zu lokalisieren. Mit dem Adressübersetzungsorakel können wir auf 64-Bit-Linux-Systemen aus unprivilegierten Benutzerprogrammen und sogar in einer virtuellen Amazon EC2-Maschine, virtuelle Adressen in physikalische Adressen übersetzen.

Wir bauen drei Angriffe aus, die diese Primitiven ausnutzen. Unser erster Angriff erstellt ein exaktes Bild der gesamten Seitentabellenhierarchie eines Prozesses und hebelt damit alle ASLR Varianten aus. Bei unserem zweiten Angriff lösen wir virtuelle auf physikalische Adressen auf, um dann SMAP auf 64-Bit-Linux-Systemen mithilfe von Angriffen im Stil von ret2dir zu umgehen. Auf der Grundlage beider Orakel haben wir demonstriert, wie Kernel ASLR unter Windows 10 besiegt werden kann – eine Grundlage für ROP-Angriffe auf Kernel- und Treiber-Binärcode. Als Gegenmaßnahme habe ich schließlich eine neue Form von Kernelisolierung vorgeschlagen. Dieser Vorschlag ist mittlerweile auch bekannt durch seine praktischen Implementationen, z.B. KAISER [Gr17] oder KPTI, die den Adressraum von Programmen vom Adressraum des Betriebssystemkernels trennt. Diese Gegenmaßnahme erfordert nur wenige – jedoch tiefgreifende – Änderungen in Betriebssystemkerneln. Die Leistungseinbuße auf aktueller Hardware und bei realistischen Anwendungsszenarien liegt zwischen 0,06% und 5,09%. Mittlerweile haben alle verbreiteten Betriebssysteme meinen Vorschlag aufgenommen und Varianten dieser Sicherheitsmaßnahme implementiert. Unsere Ergebnisse wurden auf der CCS 2016 Konferenz [Gr16a] in Zusammenarbeit mit Clémentine Maurice, Anders Fogh, Moritz Lipp und Stefan Mangard veröffentlicht.

## 4 Schlussfolgerungen

Aus meiner Dissertation und den Publikationen, die sie umfasst, können wir Rückschlüsse auf vier verschiedene Achsen ziehen.

Erstens können Mikroarchitekturangriffe weitgehend automatisiert werden. Die Automatisierung macht es deutlich einfacher, Mikroarchitekturangriffe durchzuführen. Sie ermöglicht aber auch eine Großzahl von Angriffen mit geringem zusätzlichem Aufwand.

Zweitens: Unbekannte und neuartige Seitenkanäle existieren sehr wahrscheinlich und werden sehr wahrscheinlich auch gefunden.[4] Wir haben bereits gezeigt, dass moderne Mikroar-

chitekturen mehrere zuvor unbekannte Seitenkanäle haben, wie beispielweise der `clflush`-Anweisung [GMWM16], den Arbeitsspeicher [Pe16] oder Prefetch-Anweisungen [Gr16a].

Drittens ist es möglich, die Anforderungen bekannter Angriffe auf einen Punkt zu reduzieren und zu minimieren, an dem sie in stark eingeschränkten und Sandbox-Umgebungen durchgeführt werden können. Wir haben dies in unserer Arbeit zu Rowhammer-Angriffen in JavaScript [Gr16b] und in unserer Arbeit zu Seiten-Deduplizierungsangriffen in JavaScript [GBM15] gezeigt.

Viertens ist es eine schwierige Aufgabe, effektive und effiziente Gegenmaßnahmen zu entwickeln. Die Forschung versucht oft überambitioniert, universelle Gegenmaßnahmen gegen Mikroarchitekturangriffe zu finden und ignoriert, dass die verschiedenen Angriffe sehr unterschiedliche Anforderungen und Eigenschaften haben [GMWM16; Pe16]. Im Mittelpunkt von Mikroarchitekturangriffen steht meist ein zeitlicher oder verhaltensbedingter Unterschied, der vom Prozessorhersteller zur Optimierung der Performance angestrebt wird. Daher kann es schwierig sein, immer eine universelle Gegenmaßnahme zu finden, die die Leistung nicht beeinträchtigt, da Sicherheit und Leistung sich oft widersprechen [Gr16a].

## Literaturverzeichnis


[GBM15] D. Gruss, D. Bidner und S. Mangard. “Practical Memory Deduplication Attacks in Sandboxed JavaScript”. In: *ESORICS*. 2015.

[GMWM16] D. Gruss, C. Maurice, K. Wagner und S. Mangard. “Flush+Flush: A Fast and Stealthy Cache Attack”. In: *DIMVA*. 2016.

[Gr16a] D. Gruss, C. Maurice, A. Fogh, M. Lipp und S. Mangard. “Prefetch Side-Channel Attacks: Bypassing SMAP and Kernel ASLR”. In: *CCS*. 2016.

[Gr16b] D. Gruss, C. Maurice und S. Mangard. “Rowhammer.js: A Remote Software-Induced Fault Attack in JavaScript”. In: *DIMVA*. 2016.

[Gr17] D. Gruss, M. Lipp, M. Schwarz, R. Fellner, C. Maurice und S. Mangard. “KASLR is Dead: Long Live KASLR”. In: *ESSoS*. 2017.

[Gru17] D. Gruss. “Software-based Microarchitectural Attacks”. Diss. Graz University of Technology, 2017.

[GSM15] D. Gruss, R. Spreitzer und S. Mangard. “Cache Template Attacks: Automating Attacks on Inclusive Last-Level Caches”. In: *USENIX Security Symposium*. 2015.

[HWH13] R. Hund, C. Willems und T. Holz. “Practical Timing Side Channel Attacks against Kernel Space ASLR”. In: *S&P*. 2013.

[Ki2014] Y. Kim, R. Daly, J. Kim, C. Fallin, J. H. Lee, D. Lee, C. Wilkerson, K. Lai und O. Mutlu. “Flipping bits in memory without accessing them: An experimental study of DRAM disturbance errors”. In: *ISCA*. 2014.


[4] Diese Vorhersage aus meiner Dissertation hat sich in einem unvorhergesehenen Ausmaß bewahrheitet, und zwar mit Meltdown [Li18] und Spectre [Ko19].

[Ko19] P. Kocher, J. Horn, A. Fogh, D. Genkin, D. Gruss, W. Haas, M. Hamburg, M. Lipp, S. Mangard, T. Prescher, M. Schwarz und Y. Yarom. “Spectre Attacks: Exploiting Speculative Execution”. In: *S&P*. 2019.

[Koc96] P. C. Kocher. “Timing Attacks on Implementations of Diffe-Hellman, RSA, DSS, and Other Systems”. In: *Crypto*. 1996.

[Li16] M. Lipp, D. Gruss, R. Spreitzer, C. Maurice und S. Mangard. “ARMageddon: Cache Attacks on Mobile Devices”. In: *USENIX Security Symposium*. 2016.

[Li18] M. Lipp, M. Schwarz, D. Gruss, T. Prescher, W. Haas, S. Mangard, P. Kocher, D. Genkin, Y. Yarom und M. Hamburg. “Meltdown: Reading Kernel Memory from User Space”. In: *USENIX Security Symposium*. 2018.

[OST06] D. A. Osvik, A. Shamir und E. Tromer. “Cache Attacks and Countermeasures: the Case of AES”. In: *CT-RSA*. 2006.

[Pe16] P. Pessl, D. Gruss, C. Maurice, M. Schwarz und S. Mangard. “DRAMA: Exploiting DRAM Addressing for Cross-CPU Attacks”. In: *USENIX Security Symposium*. 2016.

[SD15] M. Seaborn und T. Dullien. “Exploiting the DRAM rowhammer bug to gain kernel privileges”. In: *Black Hat Briefings*. 2015.

[SIYA11] K. Suzaki, K. Iijima, T. Yagi und C. Artho. “Memory Deduplication as a Threat to the Guest OS”. In: *EuroSec*. 2011.

[YF14] Y. Yarom und K. Falkner. “Flush+Reload: a High Resolution, Low Noise, L3 Cache Side-Channel Attack”. In: *USENIX Security Symposium*. 2014.

[ZJRR14] Y. Zhang, A. Juels, M. K. Reiter und T. Ristenpart. “Cross-Tenant Side-Channel Attacks in PaaS Clouds”. In: *CCS*. 2014.

**Daniel Gruss**, geboren am 16. September 1986 in Brühl, Deutschland. Er begann sein Informatikstudium 2008 an der Technischen Universität Graz und promovierte 2017 mit Auszeichnung. Er erhielt eine Auszeichnung für die beste Bachelorarbeit am Institut für Angewandte Informationsverarbeitung und Kommunikation der Technischen Universität Graz. Er hält regelmäßig Vorträge bei akademischen und industriellen IT-Sicherheitskonferenzen. Mit durchschnittlich drei Top-Tier-Publikationen pro Jahr zählt Daniel zu den produktivsten System-Security-Forschern. Seine Forschung ist nicht nur in der Wissenschaft, sondern auch in der Industrie bekannt, wie zum Beispiel die erste Rowhammer-Attacke aus Sandbox-JavaScript oder die Meltdown- und Spectre-Attacken, die enorme Auswirkungen auf die gesamte digitale Welt hatten.

Steffen Hölldobler et. al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2018 91

# Spontane Sicherheitsprüfung mittels individualisierter Programmzertifizierung oder Programmrestrukturierung[1]

Marie-Christine Jakobs[2]

**Abstract:** Korrekt funktionierende Software gewinnt immer mehr an Bedeutung. Im Vergleich zu früher ist es heutzutage schwieriger einzuschätzen, wie gut eine Software funktioniert. Dies liegt unter anderem daran, dass Endnutzer häufiger Software unbekannter Hersteller installieren. Endnutzer sollten sich also aktiv von der Softwarekorrektheit überzeugen, zum Beispiel in Form einer spontanen Sicherheitsprüfung. Übliche Verifikationstechniken zur Korrektheitsprüfung kommen für Endnutzer, in der Regel Laien, nicht in Frage. Die zentrale Frage ist daher, wie man einem Laien eine solche spontane Sicherheitsprüfung ermöglicht. Die Antwort der Dissertation sind einfache, automatische und generelle Verfahren zur Sicherheitsprüfung. In der Dissertation werden verschiedene Verfahren vorgeschlagen und sowohl theoretisch als auch praktisch untersucht. Die vorgeschlagenen Verfahren lassen sich in zwei Forschungsrichtungen einsortieren, nämlich in die Gruppe der Proof-Carrying Code Verfahren bzw. in die Gruppe des alternativen Programs from Proofs Verfahren. Einige Verfahren kombinieren beide Forschungsrichtungen.

## 1 Motivation

Im Zeitalter von mobilen Endgeräten und „smarten“ Geräten ist es Gang und Gäbe die Funktionen der Geräte mittels Softwareapplikationen aus dem Internet zu erweitern. (Unbeabsichtigte) Fehler, die in fast jeder Softwareapplikation existieren, können unsere Sicherheit bedrohen, zumindest aber den Betrieb eines unserer Geräte stören. Da wir uns im Alltag immer mehr auf unsere Geräte verlassen, kann ein solcher Fehler einen dramatischen Einfluss auf unser Leben haben.

Früher haben wir überwiegend Software von wenigen, namhaften Herstellern installiert, bei denen wir aufgrund von Erfahrung einschätzen konnten, wie gut die Qualität ihrer Produkte ist, das heißt, wie fehleranfällig ihre Produkte sind. Mit dem Aufkommen der mobilen Endgeräte hat sich unser Verhalten geändert. Wir installieren immer häufiger Applikationen von vielen verschiedenen, uns unbekannten Hersteller. Eine Abschätzung der Fehleranfälligkeit von Applikationen wird deutlich schwieriger. Mehr denn je sollten wir Nutzer uns vor der Installation selbst aktiv davon überzeugen, dass eine Softwareapplikation unseren Qualitätsansprüchen genügt.

Verifikation ist ein geeignetes Mittel der Qualitätsprüfung, sofern man Experte ist. Tatsächlich sind die meisten Nutzer aber Verifikationslaien und scheitern bereits daran, das richtige Verfahren für ihr Verifikationsproblem aus der Vielzahl der Verfahren zu wählen.

[1] Englischer Titel der Dissertation: “On-the-fly Safety Checking – Customizing Program Certification and Program Restructuring”

[2] Software and Computational Systems Lab, LMU Munich, jakobs@sosy.ifi.lmu.de

Es wird also ein möglichst einfaches, generelles und automatisches Verfahren benötigt, dass die spontane Qualitätsprüfung zuverlässig für den Nutzer übernimmt. Da wir uns im Folgenden insbesondere für Sicherheitseigenschaften (im Sinne von „safety“) interessieren, nennen wir die Qualitätsprüfung von nun an Sicherheitsprüfung.

Bereits um die Jahrtausendwende herum hat sich die Verifikationsgemeinschaft mit einer ähnlichen Fragestellung beschäftigt. Die Lösung für die spontane Sicherheitsprüfung war das von Necula vorgeschlagene Proof-Carrying Code Protokoll [Ne97]. Die Idee des Protokolls ist, dass der Softwarehersteller die Verifikation der Software übernimmt und dem Nutzer seine aus der Verifikation gewonnen Beweisinformationen zur Verfügung stellt. Die spontane Sicherheitsprüfung des Nutzers umfasst dann nur noch die einfachere Beweisprüfung. Die Gemeinschaft hat eine Vielzahl von speziellen Proof-Carrying Code Verfahren für diverse Analysen und Eigenschaftsklassen entwickelt. Kürzlich wurde sogar ein alternatives Protokoll, genannt Programs from Proofs [WSW13], vorgeschlagen. Allerdings habe alle Ansätze in der Literatur einen entscheidenden Nachteil. Sie sind nicht allgemein anwendbar oder nicht automatisch.

Ziel der Dissertation [Ja17] ist es daher generelle, automatische Verfahren zur spontanen Sicherheitsprüfung zu entwickeln. Die Verfahren bauen auf dem Proof-Carrying Code Protokoll oder dem Programs from Proofs Protokoll auf. Es werden auch Verfahren entwickelt, die eine Kombination beider Protokolle nutzen. Außerdem werden die Verfahren nicht nur entwickelt, sondern auch hinsichtlich ihrer Nützlichkeit für eine spontane Sicherheitsprüfung untersucht. Zur Nützlichkeit gehören sowohl theoretische Betrachtungen, zum Beispiel ist das Verfahren zuverlässig, als auch praktische Fragestellungen, zum Beispiel wie effizient ist die spontane Sicherheitsprüfung.

Der nächste Abschnitt gibt einen genaueren Überblick über die Anforderungen an die entwickelten Verfahren und eine schematische Darstellung ihrer allgemeinen Funktionsweise. Die Details der verschiedenen Verfahren sowie ihre Eigenschaften werden im danach folgenden Abschnitt beleuchtet. Daran anschließend werden die entwickelten Verfahren verglichen. Zum Schluss wird ein Fazit gezogen.

## 2 Überblick

In der Dissertation werden verschiedene Verfahren zur spontanen Sicherheitsprüfung entwickelt und evaluiert. Die Verfahren gehören zu einem von zwei Forschungsansätzen (Proof-Carrying Code [Ne97], Programs from Proofs [WSW13]) oder deren Kombination. Nichtsdestotrotz verfolgen alle entwickelten Verfahren die gleichen folgenden Ziele.

**Stichhaltigkeit** Bestätigt die spontane Sicherheitsprüfung eine Programmeigenschaft, so gilt diese.

**Vollständigkeit** Eine korrekte Vorbereitung der spontanen Sicherheitsprüfung garantiert ihren Erfolg.

**Automatismus** Mit Ausnahme einer anfänglichen Konfiguration der Vorbereitungsphase verläuft die Sicherheitsprüfung vollautomatisch ohne menschliche Intervention.

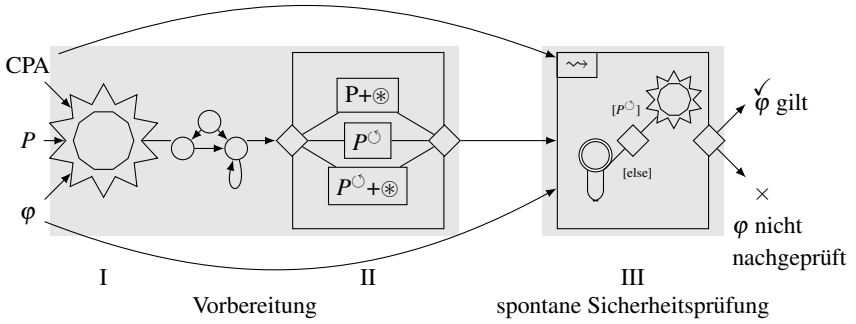



Abb. 1: Schematische Darstellung der Funktionsweise der spontanen Sicherheitsprüfungen

**Generalität** Die spontane Sicherheitsprüfung unterstützt verschiedene Eigenschaften.

**Effizienz** Die spontane Sicherheitsprüfung ist schneller und benötigt weniger Speicher als eine normale Verifikation.

Stichhaltigkeit und Vollständigkeit der Verfahren werden formal bewiesen. In wie weit die Verfahren Automatismus und Generalität unterstützen wird diskutiert und die Effizienz wird experimentell evaluiert.

Eine weitere Gemeinsamkeit der Verfahren ist ihre abstrakte Funktionsweise. Trotz all ihrer Detailunterschiede folgen alle entwickelten Verfahren dem gleichen Schema, das in Abb. 1 skizziert ist. Zuerst muss die spontane Sicherheitsprüfung vorbereitet werden (linker, grauer Kasten in Abb. 1). Die Vorbereitung wird in aller Regel nicht vom spontan Prüfenden sondern vom Programmhersteller, -anbieter, etc. durchgeführt und geschieht geplant vorab.

Der erste Vorbereitungsschritt verifiziert, dass das Programm $P$ die Sicherheitseigenschaft $\varphi$ erfüllt. Die Verifikation ist für alle Verfahren gleich. Um die Generalität zu gewährleisten, nutzen wir eine parametrisierte Verifikation. Diese bekommt neben dem Programm auch die zu prüfende Eigenschaft und eine zur Eigenschaft passenden Analysekonfiguration. Die zu prüfenden „safety“ Eigenschaften werden in sogenannten Eigenschaftsautomaten kodiert. Unter anderem können Eigenschaftsautomaten die Nichterreichbarkeit von Programmzuständen, Invarianten und Protokolle spezifizieren. Allgemein gesehen kodieren sie die Menge der verbotenen Ausführungspfade. Für die Analysekonfiguration verwenden wir das Prinzip der konfigurierbaren Programmanalyse (CPA) [BHT07]. Eine konfigurierbare Programmanalyse beschreibt eine abstrakte Programmsemantik und Operatoren zur Steuerung einer abstrakten Erreichbarkeitsanalyse. Hauptsächlich gesteuert durch die konfigurierbare Programmanalyse führt die parametrisierte Verifikation genauso eine abstrakte Erreichbarkeitsanalyse durch. Das Ergebnis einer erfolgreichen Verifikation ist ein abstrakter Erreichbarkeitsgraph, eine abstrakte Beschreibung des Zustandsraums des analysierten Programms.

Im zweiten Vorbereitungsschritt werden Information aus dem Erreichbarkeitsgraph extrahiert und für die spontane Sicherheitsprüfung zur Verfügung gestellt. Die Vorgehensweise ist dabei abhängig vom Forschungsansatz. Die Verfahren im Bereich Proof-Carrying Code geben dem Programm ein separates *Zertifikat* ($\circledast$) mit. Das Zertifikat speichert Teile des Erreichbarkeitsgraphen. Verfahren im Bereich Programs from Proofs kodieren die extrahierten Informationen direkt in einem *transformierten Programm* $P^{\circlearrowleft}$. Das transformierte Programm ist eine Restrukturierung des Originalprogramms $P$, in der unmögliche Pfade gelöscht und Ausführungspfade syntaktisch voneinander getrennt wurden. In der Kombination wird sowohl die Programmtransformation aus dem Bereich Programs from Proofs angewendet, als auch ein separates Zertifikat für das transformierte Programm erstellt.

Der dritte und letzte Schritt ist die spontane Sicherheitsprüfung. Neben der zu prüfenden Eigenschaft und dem im Schritt II generierten Artefakt bekommt die spontane Sicherheitsprüfung auch die ursprüngliche Analysekonfiguration. Diese wird benötigt, um die spontane Sicherheitsprüfung automatisch an die zu prüfende Eigenschaft anzupassen. Dafür wird aus der ursprünglichen Konfiguration automatisch eine Konfiguration für die spontane Sicherheitsprüfung abgeleitet ($\rightsquigarrow$), die auf die Prüfung der Eigenschaft mittels des Artefaktes abgestimmt ist. Die Verfahren im Bereich Proof-Carrying Code und der Kombination inspizieren das Zertifikat und prüfen, ob es valide bezeugt, dass das Programm im Artefakt die zu prüfenden Eigenschaft erfüllt. Dagegen verifizieren Verfahren im Bereich Programs from Proofs das transformierte Programm bezüglich der zu prüfenden Eigenschaft unter Verwendung des gleichen Verfahrens wie im Schritt I der Vorbereitung. Details folgen im nächsten Abschnitt, in dem wir einen genaueren Blick auf die konkret entwickelten Verfahren werfen.

## 3 Verfahren zur spontanen Sicherheitsprüfung

Wie bereits im Überblick skizziert, verfolgt die Dissertation drei grundsätzliche Techniken zur Realisierung der spontanen Sicherheitsprüfung. Die folgenden Abschnitte geben einen Einblick in die entwickelten Verfahren der jeweiligen Technik.

### 3.1 Konfigurierbare Softwarezertifizierung

Die Verfahren zur konfigurierbaren Softwarezertifizierung folgen der Proof-Carrying Code Idee [Ne97]. Das bedeutet, dass mit dem Programm Teile des Korrektheitsbeweises in Form eines Zertifikats mitgegeben werden. Die Teile des Korrektheitsbeweises ermöglichen die spontanen Sicherheitsprüfung, welche den Beweis effizient nachprüft, das heißt, ihn nachvollzieht und gegebenenfalls in Teilen nachberechnet. In unserem Fall liegt der Beweis in Form des abstrakten Erreichbarkeitsgraphens vor. Seine zentrale Eigenschaft ist eine sichere Überapproximation der erreichbaren Programmzustände, das heißt, die Überapproximation verletzt die zu prüfende Eigenschaft nicht. Im Prinzip ist die Überapproximation vollständig durch die Knoten des Graphens beschrieben. Denn die Knoten repräsentieren die abstrakten Zustände, das heißt, Mengen von realen Programmzuständen.

Daher speichern die Zertifikate der konfigurierbare Softwarezertifizierung lediglich Knoten. Für die konfigurierbare Softwarezertifizierung wurden Verfahren mit verschiedenen Zertifikatstypen entworfen. Aufbauend auf einem simplen Basisverfahren wurden zwei orthogonale Optimierungen, die unterschiedliche Optimierungsziele verfolgen, entwickelt und letztendlich in einem Verfahren integriert. Die folgende Aufstellung gibt einen Überblick über die betrachteten Zertifikatstypen.

1. Basis: Speicherung aller Knoten
2. Optimierungen
   a) Reduktion: Speicherung einer Teilmenge der Knoten
   b) Aufteilung: Speicherung einer Partition der Menge der Knoten
3. Integration: Speicherung einer Partition einer Teilmenge der Knoten

Ziel beider Optimierungen ist es, die Zeit der Zertifikatsprüfung zu verringern. Die Reduktionsoptimierung reduziert die Größe des Zertifikats und somit die Einlesezeit des Zertifikates. Gleichzeitig verhindert sie unnötige Nachprüfungen. Welche Knoten mindestens gespeichert werden müssen hängt von der Struktur des Graphens und der Analyseart ab. Die Details befinden sich in Kapitel 4.1 der Dissertation. Die Aufteilungsoptimierung reduziert die Prüfungszeit, indem sie Einlesen und Nachprüfung parallelisiert. Die Integration kombiniert die vorherigen Optimierungen. Zusätzlich kann man mit ihr alle vorherigen Zertifikate beschreiben, wenn auch nicht in syntaktisch identischer Form, da die Partitionsgröße und die zu speichernde Teilmenge der Knoten prinzipiell eingestellt werden können.

Die Nachprüfung des Beweises mit Hilfe des Zertifikates ist im Wesentlichen eine Überprüfung, ob das Zertifikat eine sichere Überapproximation darstellt oder sich zu einer sicheren Überapproximation erweitern lässt. Dafür werden die abstrakte Semantik der initialen konfigurierbaren Programmanalyse verwendet und Teile ihrer Operatoren im Prinzip wiederverwendet. Die konkrete Überprüfung unterscheidet sich für die verschiedenen Zertifikatstypen. Allerdings, bleibt die Grundidee für alle Verfahren die gleiche:

- Nachberechnen der abstrakten Nachfolgezustände aller gespeicherten und explorierten Zustände.
- Prüfen, ob die Nachfolgezustände von gespeicherten Zuständen überdeckt werden. Nicht überdeckte Nachfolgezustände werden in die Menge der explorierten Zustände aufgenommen.
- Prüfen, ob die finale Abstraktion sicher ist.
- Scheitern, sobald zu viele Zustände exploriert wurden.[3]

Betrachten wir die fünf Ziele, so ergibt sich folgendes Bild. Die Stichhaltigkeit und Vollständigkeit werden in der Dissertation für alle Zertifikatstypen und -verfahren formal bewiesen. Die Automatisierung funktioniert für viele Analysen und alle Eigenschaften. Für

[3] Dieses Abbruchkriterium ist notwendig, um die Terminierung der Überprüfung und somit die Vollständigkeit des Verfahrens zu garantieren.

einige Analysen muss ein verwendeter Analyseoperator manuell präzisiert werden. Somit sind die Verfahren generell.

Die Effizienz wurde mit 20 Analysekonfigurationen und einer Teilmenge einer großen Softwareverifikationsbenchmark evaluiert. Die Ergebnisse der Evaluation sind wie folgt. Die Effizienz des Ansatzes hängt stark von der Analyse ab, die durch Konfiguration, Eigenschaft und Programm bestimmt wird. In der Regel sind die Optimierungen effizienter als der Basisansatz. Die besten Optimierungen sind die Reduktionsoptimierung sowie die Integration. Zwischen den beiden gibt es keinen klaren Sieger. Im Vergleich mit ähnlichen Verfahren zeigt sich, dass diese höchstens einen kleinen Vorteil auf Analysetypen haben, für die sie ursprünglich entwickelt wurden. Im Allgemeinen kann die jeweils beste Optimierung mit den anderen Verfahren mithalten, obwohl ihre Zertifikate deutlich größer als das Programm selbst und auch teils größer als manche Zertifikate konkurrierender Ansätze sind.

### 3.2 Generisches Programs from Proofs

Ziel des Programs from Proofs Verfahrens ist es, simple und effiziente Datenflussanalysen für die spontanen Sicherheitsprüfungen zu verwenden. Allerdings scheitern Datenflussanalysen alleine häufig daran, die gewünschte Programmeigenschaft zu zeigen, da sie lediglich fluss- aber nicht pfadsensitiv sind. Schauen wir uns zunächst das linke Programm in Abb. 2 an. Um zu zeigen, dass vor dem Fahrversuch, das Fahrzeug gestartet wird, muss die Analyse wissen, dass die beiden if-Blöcke über die Bedingung $p$ gekoppelt sind. Betrachten wir nun das rechte Programm in Abb. 2. Da das rechte Programm die beiden if-Blöcke des linken Programms zu einem Block zusammenfasst, ist ein Wissen über $p$ unnötig. Die Programmstruktur alleine genügt, um die beschriebene Eigenschaft zu zeigen. Die Restrukturierung des linken Programms ermöglicht also eine simple und effiziente Eigenschaftsanalyse. Analysen, die die gewünschte Programmeigenschaft nachweisen können, machen häufig implizit eine solche Programmrestrukturierung und ihre Restrukturierung findet sich im abstrakten Erreichbarkeitsgraph wieder. Dies macht sich das Programs from Proofs Verfahren, welches zum ersten Mal von Wonisch et al. [WSW13] vorgeschlagen wurde, zu Nutze. Beginnend mit einer kombinierten Analyse, die die effiziente Zielanalyse inkorporiert, aber dennoch mächtiger ist, wird das Programm verifiziert. Um die Eigenschaft mit der Zielanalyse durchführen zu können, bedient sich der Ansatz eines Tricks. Die im Erreichbarkeitsgraphen kodierte Restrukturierung wird ins Programm übertragen. Konkret wird der Erreichbarkeitsgraph, insbesondere seine Struktur, in ein Programm übersetzt. Wonisch et al. [WSW13] haben dieses Prinzip lediglich auf eine spezielle Analyse und Eigenschaftsklasse angewendet. In der Dissertation wird das Prinzip generalisiert.

```
if(p)          if(p)
  start();       start();
if(p)            fahr();
  fahr();
```

Abb. 2: Programm mit Restrukturierung

Im Gegensatz zur vorherigen konfigurierbaren Softwarezertifizierung verhindert die Programmrestrukturierung im Verfahren, dass dieses Verfahren mit allen beliebigen Eigenschaften und Analysen funktioniert. Eigenschaften dürfen sich nicht direkt auf den Pro-

grammzähler („program counter“) beziehen. Ebenso benötigen wir für die anfängliche konfigurierbare Programmanalyse eine bestimmte strukturelle Beschaffenheit. So besteht diese Analyse aus zwei Komponenten: dem Eigenschaftsprüfer und der Hilfsanalyse. Der Eigenschaftsprüfer alleine prüft die Eigenschaft. Er ist flusssensitiv und resistent gegenüber Programmrestrukturierung. Seine Datenflussvariante wird später für die spontane Sicherheitsprüfung verwendet. Die Hilfsanalyse darf unmögliche Pfade ausschließen und Programmausführungen voneinander separieren. Sie beeinflusst den Eigenschaftsprüfer aber nie direkt. Die Details können in der Dissertation [Ja17, S. 150ff] nachgelesen werden.

Die Generalität des Verfahrens wurde im vorherigen Absatz bereits beleuchtet. Kommen wir nun zu den vier restlichen Zielen. Da das Verfahren für die spontane Sicherheitsprüfung das Analyseverfahren aus der Vorbereitung nutzt, ergibt sich die Stichhaltigkeit aus diesem Analyseverfahren, welches selbst stichhaltig ist. Der Nachweis der Vollständigkeit gliedert sich in zwei Teile: a) Vollständigkeit unter der Annahme der Terminierung der Analyse und b) Terminierung. Teilaspekt a) wurde formal nachgewiesen. Der Nachweis für Teilaspekt b) gelang für verschiedene Typen von Analysen, zum Beispiel Modelchecking, und insbesondere für alle praktisch relevanten Analysen, allerdings nicht allgemein. Der Automatismus ergibt sich aus der Konstruktion des Verfahrens.

Für die Evaluation der Effizienz wurden 3 Hilfsanalysen mit 12 Eigenschaftsprüfern kombiniert, um mit ihnen das Program from Proofs Verfahren mit verschiedenen Eigenschaften auf Programmen unterschiedlicher Softwareverifikationsbenchmarks auszuführen. Neben dem Vergleich mit der normalen Verifikation, der Verifikation in der Vorbereitung, wurde die spontane Sicherheitsprüfung auch mit der Prüfung in der konfigurierbaren Softwarezertifizierung verglichen. Dabei stellte sich heraus, dass die spontane Sicherheitsprüfung des Programs from Proofs Verfahrens meist effizienter ist als die normale Verifikation und mit der konfigurierbaren Softwarezertifizierung mindestens mithalten kann. Zusätzlich ergab die Evaluation, dass die transformierten Programme häufig nicht allzu viel größer sind als das Originalprogramm, meist höchstens 5-mal so groß, und nicht selten auch kleiner.

Neben den eben betrachteten fünf Zielen ergeben sich aufgrund der Programmtransformation weitere Fragen. Zentral für die Anwendung des Ansatzes ist die Frage, ob das Originalprogramm und das transformierte Programm äquivalent sind. In der Dissertation wurde gezeigt, dass die beiden Programm sich äquivalent modulo Programmzähler verhalten, also sich die möglichen Ausführungspfade der Programm sich nur in den Werten des Programmzählers unterscheiden. Praktisch verhalten sie sich also äquivalent. Zusätzlich wurde untersucht welche Eigenschaften des Originalprogramms auf dem transformierten Programm erhalten bleiben beziehungsweise beweisbar bleiben (siehe [Ja17, S. 171f, 186ff]).

### 3.3 Konfigurierbare Softwarezertifizierung für Programs from Proofs

Ziel der Kombination ist es, von den Stärken der beiden vorherigen Ansätze zu profitieren, insbesondere soll das Terminierungsproblem des Programs from Proofs Ansatzes überwunden werden. Gleichzeitig ermöglicht eine Kombination eine noch einfachere und gegebenenfalls effizientere, spontane Prüfung.

Die naive Kombination führt die beiden Ansätze sequentiell hintereinander aus. Die Vorbereitung beginnt wie die Programs from Proofs Vorbereitung. Sie hört dort aber nicht auf, sondern setzt mit der spontanen Sicherheitsprüfung der Program from Proofs Technik fort, die gleichzeitig Schritt I der konfigurierbaren Softwarezertifizierung darstellt. Danach endet die Vorbereitung mit Schritt II der konfigurierbaren Softwarezertifizierung. Diese naive Lösung verschiebt daher lediglich das Terminierungsproblem anstatt es zu lösen. Zusätzlich wird in der Vorbereitung das Programm im Prinzip doppelt verifiziert, also unnötig viel Aufwand betrieben. Eine sinnvolle Kombination muss daher ein Zertifikat für das transformierte Programm aus dem Analyseergebnis vom Originalprogramm berechnen.

In der Dissertation werden zwei Varianten aufgezeigt, ein Zertifikat für das transformierte Programm zu berechnen:

- Transformation des abstrakten Erreichbarkeitsgraphens und Zertifikatsgenerierung aus dem transformierten Graphen
- Erstellung eines Zertifikats für das Originalprogramm und Transformation des Zertifikats auf das transformierte Programm

In beiden Fällen wird die Struktur des Graphens beziehungsweise des Zertifikates in der Transformation erhalten. Lediglich die abstrakten Zustände werden angepasst, indem die Informationen der Hilfsanalyse entfernt werden und die Programmzähler („program counter“) entsprechend dem Wissen über die Programmtransformation angepasst werden. Ebenso funktionieren beide Varianten mit allen Zertifikatsarten der konfigurierbaren Softwarezertifizierung. Die beiden Verfahren unterscheiden sich nur in zwei Aspekten. Die Transformation des abstrakten Erreichbarkeitsgraphens braucht mehr Speicher. Im Gegensatz erhält man bei der Transformation des Zertifikats für manche Zertifikatstypen teilweise größere Zertifikate.

Aufgrund der Eigenschaften der konfigurierbaren Softwarezertifizierung und des Programs from Proofs Verfahrens sind die Kombinationsverfahren selber stichhaltig, vollständig und automatisch, sofern die erzeugten Zertifikate von der konfigurierbaren Softwarezertifizierung unter Betrachtung des transformierten Programms erstellt werden könnten. Diese Eigenschaft der erstellten Zertifikate wird in der Dissertation formal für beide Varianten bewiesen. Weil die Kombination die Generalität von dem Program from Proofs Ansatz erbt, bleibt lediglich die Frage der Effizienz.

Anstatt die Kombination mit der normalen Verifikation zu vergleichen, haben wir die spontane Sicherheitsprüfung der Kombination mit der Program from Proofs Technik verglichen. Wir wollten also primär wissen, ob wir die Programs from Proofs Technik verbessern konnten. Das Ergebnis ist, dass dies eher selten möglich ist, da die spontane Sicherheitsprüfung der Programs from Proofs Technik bereits sehr effizient ist.

## 4 Vergleich der Verfahren

Wir vergleichen die entwickelten Verfahren zunächst anhand der fünf Ziele. Alle Verfahren sind stichhaltig. Die Kombination ist sowohl automatisch als auch vollständig. Im Gegensatz dazu sind die Verfahren der konfigurierbaren Softwarezertifizierung nicht immer automatisch und für die Program from Proofs Technik konnte die Vollständigkeit nicht für alle Analysen gezeigt werden. Die konfigurierbare Softwarezertifizierung ist am allgemeinsten. Sie funktioniert für alle konfigurierbaren Analysen und Eigenschaften. Der Programs from Proofs Ansatz und die Kombination funktionieren lediglich auf einer Teilmenge der Analysen und Eigenschaftsklassen. Diese Teilmenge ist für beide gleich. Bezüglich der Effizienz gilt Folgendes. Auch wenn die Verfahren mit existierenden Verfahren mithalten können, so ist das Program from Proofs Verfahren im direkten Vergleich mit der konfigurierbaren Softwarezertifizierung beziehungsweise der Kombination etwas effizienter.

Kommen wir nun zu zwei weiteren Kriterien. Als erstes betrachten wir die sogenannte „Trusted Computing Base“, das heißt, die Komponenten, deren Implementierung man trauen muss, um dem Ergebnis der spontanen Sicherheitsprüfung zu glauben. Die kleinste „Trusted Computing Base“ hat die Kombination. Der Vergleich der anderen beiden Verfahren gestaltet sich schwieriger. Zum einen ist die Frage, ob die Inspizierung des Zertifikats schwieriger zu implementieren ist als eine Erreichbarkeitsanalyse. Aufgrund der Ähnlichkeit, der Algorithmen ist ihre Implementierung ähnlich schwer. Zum anderen stellt sich die Frage, was ist fehleranfälliger die Hilfsanalyse oder die Kombination von abstrakten Zuständen. Weil die Hilfsanlysen häufig externe Tools nutzen, sind sie, so denken wir, fehleranfälliger. Es scheint so, dass die Program from Proofs Technik einen kleinen Vorteil gegenüber der konfigurierbaren Softwarezertifizierung hat. Als zweites gucken wir uns den Festplattenspeicherbedarf an. Dort ist die Kombination am ineffizientesten. Der Vergleich der anderen beiden Ansätze ergibt, dass die Zertifikate der konfigurierbaren Zertifikate häufig größer sind als die Differenz zwischen der Größe des Originalprogramms und der Größe des transformierten Programms, was den zusätzlichen Speicherbedarf des Programs from Proofs Verfahrens charakterisiert.

## 5 Fazit

Obwohl alle Verfahren zumindest stichhaltig sind, so haben sie doch unterschiedliche Stärken und Schwächen. Die Kombination empfiehlt sich, wenn Terminierung ein Problem bei der Anwendung des Program from Proofs Verfahrens ist. Ansonsten bietet sich das Programs from Proofs Verfahren für alle Eigenschaften an, die nicht direkt vom Programmzähler abhängen. Falls die Eigenschaft direkt vom Programmzähler abhängt oder keine passende Kombination aus Hilfs- und Eigenschaftsanalyse gefunden werden kann, kann man auf die konfigurierbare Softwarezertifizierung zurückgreifen.

## Literaturverzeichnis

[BHT07] Beyer, Dirk; Henzinger, Thomas A.; Théoduloz, Grégory: Configurable Software Verification: Concretizing the Convergence of Model Checking and Program Analysis. In

(Damm, Werner; Hermanns, Holger, Hrsg.): Computer Aided Verification. Jgg. 4590 in LNCS, Springer, Berlin, Heidelberg, S. 504–518, 2007.

[Ja17] Jakobs, Marie-Christine: On-The-Fly Safety Checking – Customizing Program Certification and Program Restructuring. Dissertation, Universität Paderborn, 2017.

[Ne97] Necula, George C.: Proof-carrying Code. In: Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages. ACM, New York, NY, USA, S. 106–119, 1997.

[WSW13] Wonisch, Daniel; Schremmer, Alexander; Wehrheim, Heike: Programs from Proofs – A PCC Alternative. In (Sharygina, Natasha; Veith, Helmut, Hrsg.): Computer Aided Verification. Jgg. 8044 in LNCS, Springer, Berlin, Heidelberg, S. 912–927, 2013.

**Marie-Christine Jakobs** wurde am 3. März 1987 in Bühl geboren. Sie studierte von 2006 bis 2012 Informatik an der Universität Paderborn. In dieser Zeit absolvierte sie sowohl ihren Bachelor- als auch Masterabschluss mit Auszeichnung. Im Anschluss forschte sie im Sonderforschungsbereich 901 „On-the-fly Computing“. Gleichzeitig promovierte sie an der Universität Paderborn in der Gruppe von Prof. Dr. Heike Wehrheim. Ihre Promotion schloss sie im Mai 2017 mit Auszeichnung ab. Seit Oktober 2017 ist sie Postdoktorandin an der LMU München am Lehrstuhl von Prof. Dr. Dirk Beyer.

# Strategien zur effizienten Nutzung und Erweiterung des Messfeldes in Magnetic Particle Imaging

Christian Kaethner[1]

**Abstract:** Magnetic Particle Imaging (MPI) ist ein neuartiges Verfahren der medizinischen Bildgebung, welches sich die nichtlinearen Magnetisierungseigenschaften von magnetischen Tracer-Materialien zu Nutze macht. Eine Kombination aus hoher örtlicher und zeitlicher Auflösung bei gleichzeitigem Vorliegen einer hohen Sensitivität verspricht ein hohes Potenzial für verschiedene medizinische Applikationen. Um diesem Potenzial möglichst gerecht zu werden, sollte die Generierung der verwendeten Magnetfelder sowohl möglichst effizient im Hinblick auf die aufzubringende Leistung sein als auch eine applikationsspezifische Anpassung des abtastbaren Bereiches erlauben. Die vorliegende Arbeit befasst sich mit genau diesen beiden Aspekten dieser jungen Bildgebungsmodalität und die erreichten Ergebnisse und entwickelten Strategien bieten interessante Ansatzpunkte für zukünftige Forschungsmöglichkeiten im Bereich MPI.

## 1 Einführung

Die Promotionsarbeit „Strategien zur effizienten Nutzung und Erweiterung des Messfeldes in Magnetic Particle Imaging“ [Ka17a] ist thematisch in den Bereich der medizinischen Bildgebung und Bildverarbeitung einzuordnen und befasst sich primär mit der mathematischen Erarbeitung und informatorischen Implementierung algorithmischer Lösungsstrategien. In diesem Kontext wurden sowohl simulationsgestützte als auch messbasierte Untersuchungen vorgenommen.

Um eine thematische Einordnung der Promotionsarbeit zu ermöglichen, wird zunächst mit einer generellen Einführung in die Thematik begonnen. Anschließend werden einige der zentralen Aspekte der in der Arbeit enthaltenen Originalbeiträge übersichtsartig zusammengefasst. Die Inhalte der Abschnitte basieren dabei auf den Inhalten aus [Ka17a].

### 1.1 Thematische Einordnung

Die Verwendung bildgebender Verfahren zur Unterstützung der medizinischen Diagnostik und Therapie hat sich aufgrund immer komplexerer medizinischer Fragestellungen im Laufe der Jahre zu einem essenziellen Hilfsmittel entwickelt.

Vergleicht man bereits etablierte Bildgebungsmodalitäten, wie beispielsweise die Computertomographie, die Magnetresonanztomographie oder auch die Positronen-Emissions-Tomographie, hinsichtlich ihrer örtlichen und zeitlichen Auflösung sowie der erreichbaren

[1] Institut für Medizintechnik, Universität zu Lübeck, kaethner@imt.uni-luebeck.de

Sensitivität, zeigt sich, dass die verschiedenen Ansätze einzelne Aspekte überaus gut bedienen können, keines jedoch das volle Spektrum abdeckt.

Magnetic Particle Imaging (MPI) ist ein neuartiges Verfahren, das erstmals in [GW05] vorgestellt wurde und das Potenzial zeigt dieses Spektrum vollständig zu bedienen. MPI basiert auf der Wechselwirkung magnetischer Eisenoxid-Nanopartikel mit extern angelegten oszillierenden Magnetfeldern. Für die Signalkodierung (siehe Abb. 1) wird hierbei ausgenutzt, dass sich die Magnetisierung der Nanopartikel entsprechend eines extern angelegten Magnetfeldes nicht-linear ändert. Wird ein Magnetfeld mit zeitlich oszillierender Magnetfeldstärke appliziert, führt dies zu einer wiederholten Ummagnetisierung der Partikel. Der zeitliche Verlauf der Magnetisierung resultiert hierbei in einer Modulation der Eingangsschwingung. Diese induziert unter Einbringung elektromagnetischer Spulen eine elektromagnetische Spannung, die aufgrund des Bezugs zu den Nanopartikeln als Partikelsignal bezeichnet wird. Das entsprechende Frequenzspektrum zeigt neben der Frequenz des applizierten Anregungsfeldes auch harmonische Oberschwingungen der Anregungsfrequenz, die charakteristisch für die gewählten Nanopartikel sind.

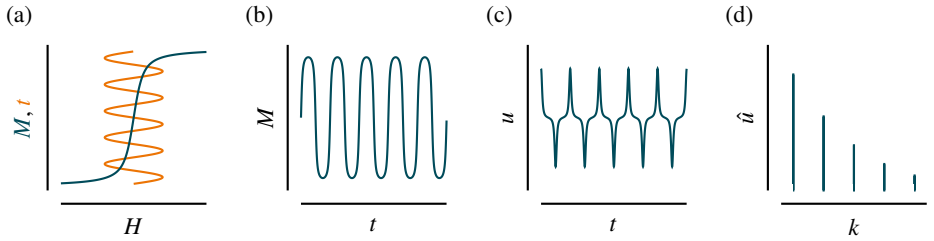



Abb. 1: Prinzip der Signalkodierung in MPI (aus [Ka17a]). (a) Das Anregungsfeld erzeugt eine zeitabhängige Magnetisierungsänderung der Partikel, (b) die einer Modulation der Eingangsschwingung entspricht. (c) Die Magnetisierungsänderung induziert ein Spannungssignal. (d) Das Frequenzspektrum des Spannungssignals zeigt neben der Anregungsfrequenz auch höhere Harmonische.

Da alle Partikel in einem Messfeld gleichermaßen auf die Änderung des Anregungsfeldes reagieren, kann für eine örtliche Unterscheidung der Partikelsignale ein zusätzliches Magnetfeld eingebracht werden. Hierbei handelt es sich um ein magnetisches Gradientenfeld, welches einen Niedrigfeldbereich aufweist, in dem die Magnetfeldstärke nahezu Null ist. Dieser sogenannte feldfreie Punkt (FFP) ermöglicht es, eine sehr genaue Lokalisation eines Partikelsignals vorzunehmen, wodurch die Verteilung der Partikel an jedem Ort und zu jedem Zeitpunkt bestimmt werden kann. Die resultierenden Signale können anschließend zur Rekonstruktion genutzt werden [Gr13] und erlauben dabei neben einer hohen örtlichen und zeitlichen Auflösung [We09], eine beeindruckende Sensitivität [Zh15]. MPI zeigt somit das Potenzial, neue Erkenntnisse im diagnostischen Bereich zu erlangen [Pa15], weswegen es zukünftig wichtig sein wird, ein Messfeld ausreichender Größe zur Verfügung zu stellen. Die Erzeugung und die Akquisition der Daten innerhalb eines solchen Messfeldes sollte dabei möglichst effizient gestaltet sein.

Aus diesem Grund befasst sich die Promotionsarbeit von Christian Kaethner mit innovativen Lösungsstrategien, die sowohl eine effiziente Nutzung eines gegebenen Messfeldes erlauben [Ka14, Er15a, Er15b, MKB16, Ka16], als auch dieses durch neuartige Konzepte

erweitern [Ka15a, Ka15b, Ka17b]. Die Ausgangssituation für die vorgestellten Ansätze war hierbei somit durch Fragestellungen aktueller Forschungsinhalte gegeben.

## 2 Leistungseffiziente Gradientenfeldgenerierung

Die leistungseffiziente Generierung eines beliebig positionierbaren FFP stellt besonders in den Randbereichen eines großen Messfeldes eine Herausforderung dar. Im Vorfeld dieser Promotionsarbeit wurde in 1D-Simulationen gezeigt, dass über eine alternative Anordnung der feldgenerierenden Spulen und eine Optimierung der elektrischen Ströme jeder einzelnen Spule eine deutliche Leistungsreduktion erreicht werden kann [KSB12]. Da hieraus jedoch nicht ohne Weiteres folgte, dass dieser Ansatz auf mehrdimensionale Szenarien erweiterbar sein würde, wurde diese Idee im Rahmen dieser Arbeit aufgegriffen und deren Machbarkeit gezeigt. Das Minimierungsproblem zur Optimierung der elektrischen Ströme konnte hierbei verallgemeinert und unabhängig von der betrachteten Dimensionalität formuliert werden.

Um das Potenzial des Ansatzes weiter zu untersuchen, wurde zunächst eine Anpassung der Spulen an die neuen Gegebenheiten des Messfeldes vorgenommen. In weiteren Schritten wurden anschließend neuartige Spulenanordnungskonzepte untersucht und verschiedene mehrdimensionale Erweiterungen vorgestellt. Ein Beispiel eines solchen Konzeptes sowie die korrespondierenden Leistungswerte sind in Abb. 2 gezeigt.

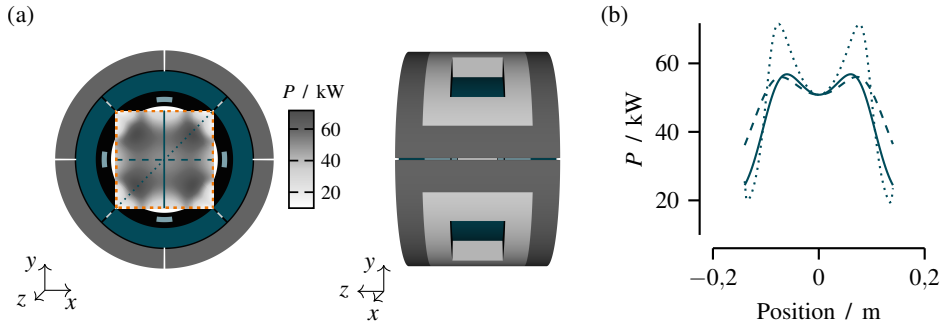



Abb. 2: Spulenanordnung unter Annahme mehrerer Schichten und radial gekrümmter Rechteckspulen (aus [Ka17a]). (a) Spulenanordnung in der Frontal- und Seitenansicht sowie Leistungsverlust im Messfeld nach Optimierung der Ströme. (b) Leistungsverlust für eine FFP-Bewegung entlang der vertikalen (——) und horizontalen Hauptachse (– –) und der Gegendiagonalen (····).

Die im Rahmen der Untersuchungen erreichte Verringerung der maximalen elektrischen Verlustleistung beträgt mehrere Zehnerpotenzen und kann somit maßgeblich zur Planung leistungseffizienterer Spulensysteme sowie der damit einhergehenden effizienteren Nutzung des gegebenen Messfeldes beitragen.

Für eine bessere Einordnung in die aktuelle Forschung sei darauf verwiesen, dass Aspekte der erreichten Ergebnisse für die Verwendung im ersten Demonstrator für die MPI-Humananwendung in Betracht gezogen wurden.

## 3 Charakterisierung von Lissajous-Kurven

Die mehrdimensionale Abtastung eines Messfeldes ist ein wichtiger Bestandteil der MPI-Bildgebung und kann über definierte Magnetfeldvariationen realisiert werden. In [Kn09] wurde gezeigt, dass hierbei die Abtastung entlang einer Lissajous-Kurve (siehe Abb. 3 (a)) besonders vorteilhaft ist. Um eine umfassende Evaluierung der Bildgebungseigenschaften solcher Kurven zu ermöglichen, wurden im Rahmen dieser Arbeit ausgewählte Charakteristika von Lissajous-Kurven, wie die Trajektoriendichte oder auch die Geschwindigkeit und Richtung der Abtastung in verschiedenen Abschnitten einer solchen Kurve, zusammengefasst und insbesondere eine Menge spezifischer Knotenpunkte eingeführt.

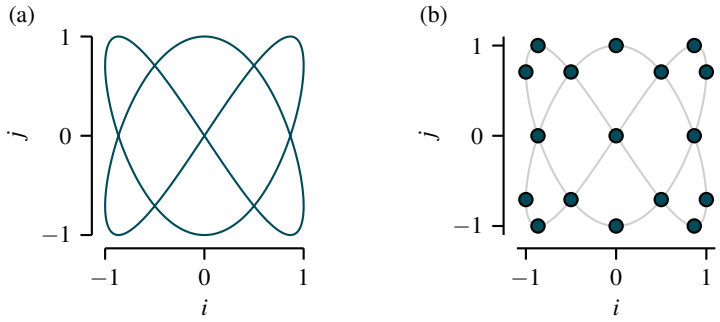



Abb. 3: (a) Beispielhafte Darstellung einer Lissajous-Kurve und (b) einer Lissajous-Kurve inklusive der korrespondierenden Lissajous-Knotenpunkte (jeweils aus [Ka17a]).

Diese sogenannten Lissajous-Knotenpunkte (siehe Abb. 3 (b)) stellen vereinfacht eine zeitlich äquidistante Abtastung der Trajektorie dar. Aufbauend auf der allgemeinen Definition der Knotenpunkte wurde sich mit verschiedenen Anwendungsmöglichkeiten befasst, die mathematischen Eigenschaften der Knotenpunkte charakterisiert und eine Verwandtschaft zu den Extremstellen von Chebyshev-Polynomen erster Art hergestellt. Im Speziellen wurde gezeigt, dass letztere eine sehr kompakte Beschreibung der Lissajous-Knotenpunkte erlauben. Die Punkte bieten zudem eine interessante Ausgangsposition für das Aufstellen eines Schemas einer polynomiellen Lagrange-Interpolation, welches in diesem Rahmen sowohl beschrieben als auch anhand numerischer Vergleichssimulationen evaluiert wurde. Es zeigte sich, dass Lissajous-Knotenpunkte für eine bivariate Interpolation ebenso gut geeignet sind, wie bereits etablierte Punktemengen, zusätzlich jedoch den Vorteil bieten einen direkten Bezug zum Abtastpfad in MPI aufzuweisen.

Die entwickelten mathematischen Konzepte der Lissajous-Knotenpunkte wurden daher im weiteren Verlauf der Promotionsarbeit im Bereich von MPI angewendet.

## 4 Systemmatrix-Akquisition an nicht-äquidistanten Gittern

Die Kalibrierung eines MPI-Systems kann über die Vermessung einer Punktprobe an verschiedenen Positionen des Systems erfolgen. Die gewählten Positionen sind dabei in der Regel gitterförmig und äquidistant verteilt. Die akquirierten Signalspektren werden an-

schließend in Form einer Systemmatrix angeordnet, die zur Bildrekonstruktion herangezogen werden kann. Aufbauend auf den Lissajous-Knotenpunkten, wurde in diesem Kontext erstmals eine nicht-äquidistante Anordnung und Reduzierung der Messpositionen untersucht (siehe Abb. 4 (a)), wodurch die inhärente Struktur der Abtasttrajektorie, einer Lissajous-Kurve, widergespiegelt wird.

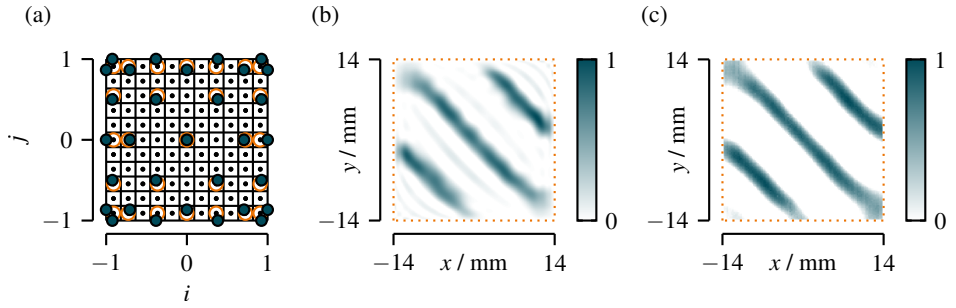



Abb. 4: Übersicht einer Systemmatrix-Akquisition an nicht-äquidistanten Gittern (aus [Ka17a]). (a) Beispielhafte Darstellung der Verteilung der Messpositionen einer Systemmatrix auf Basis von Lissajous-Knotenpunkten. (b) Rekonstruktionsergebnis unter Verwendung einer Systemmatrix wie in (a) dargestellt. (c) Rekonstruktionsergebnis auf Basis einer vollen (gitterförmigen und äquidistant verteilten) Systemmatrix-Akquisition.

Da eine Veränderung der Punktprobenpositionen eine Veränderung der Datengrundlage für die Rekonstruktion der Partikelverteilung nach sich zieht, wurde im Rahmen dieser Promotionsarbeit eine neuartige Rekonstruktionsstrategie vorgeschlagen. Die nicht-äquidistante Anordnung der Punktproben wird hierbei über eine örtliche Gewichtung basierend auf einer Voronoi-Zerlegung berücksichtigt und entsprechend in die Rekonstruktion miteinbezogen. Da die rekonstruierten Signale somit ausschließlich an den Positionen der gewählten Knotenpunkte lokalisiert sind, müssen die Signale in einem zweiten Schritt interpoliert werden, um eine Bildrepräsentation zu erhalten. Als möglicher Ansatz wurde hierzu die bereits erwähnte polynomielle Lagrange-Interpolation vorgeschlagen. Die erreichten Ergebnisse (siehe Abb. 4 (b)) weisen keine signifikanten Unterschiede zu denen einer vollen (gitterförmigen und äquidistant verteilten) Systemmatrix-Akquisition auf (siehe Abb. 4 (c)), benötigen zudem deutlich weniger Zeit und erlauben aufgrund der reduzierten Datenmenge - im untersuchten Bildgebungsszenario weniger als vier Prozent der Punktprobenpositionen einer vollen Systemmatrix-Akquisition - eine effizientere Verarbeitung.

Neben einer direkten Abbildung des verwendeten Akquisitionspfades, ist ein wesentliches Merkmal des vorgeschlagenen Akquisitionsgitters, dass die zugrundeliegende mathematische Theorie ausführlich bewiesen und einfach zu implementieren ist sowie eine schnelle und stabile Rekonstruktion der Daten ermöglicht.

# 5 Axiale Vergrößerung des Messfeldes

Die Sequenzentwicklung hinsichtlich großer Messfelder stellt in MPI eine besondere Herausforderung dar. Entscheidende Faktoren hierfür sind die Einhaltung technischer und medizinischer Sicherheitslimits sowie die damit einhergehende Limitierung der applizierten Magnetfelder und somit die physikalisch limitierte Ausdehnung des Messfeldes. Besonders interessant für die Bildgebung großer Volumen ist in diesem Zusammenhang eine axiale Erweiterung des Messfeldes über den Bildgebungsbereich eines MPI-Systems hinaus. Im Rahmen dieses Kapitels wurden daher Ansätze untersucht, die diese Thematik adressieren und sich innerhalb der genannten Magnetfeldbeschränkungen befinden.

## 5.1 Geometrische Spulenanpassung

Die asymmetrische Anordnung der feldgenerierenden Komponenten eines MPI-Systems wurde bereits im Vorfeld dieser Promotionsarbeit vorgeschlagen und erlaubt eine völlig neue Definition des Messfeldes, da alle Komponenten von einer Seite an ein Messobjekt herangeführt werden können [Sa09]. Der Zugang zu einem solchen Messobjekt ist somit nahezu unbeschränkt und erlaubt somit eine interessante Ausgangsposition für die Verwendung in interventionellen Bildgebungsszenarien. Bisherige Ansätze verwenden hierbei jedoch ausschließlich ein Konzept basierend auf kreisrunden Spulen.

Im Rahmen dieser Arbeit wurde untersucht, wie sich eine axiale Verlängerung der geometrischen Abmessungen kreisrunder Spulen (siehe Abb. 5 (a)), also eine Veränderung des Hauptachsenverhältnisses, auf die Bildgebungseigenschaften des Systems auswirken kann. Die Simulationsexperimente unter Verwendung einer asymmetrischen Spulenanordnung (siehe Abb. 5 (b)) haben ergeben, dass bei geringer Vergrößerung der Seitenverhältnisse eine bessere Rekonstruktion für einen größeren Bereich erreicht werden kann (siehe Abb. 6). Wird das Seitenverhältnis allerdings zu groß gewählt, korreliert dies mit einem Abfall in der erreichbaren Gradientenstärke in Richtung der axialen Verlängerung, was folglich zu deutlich schlechteren Rekonstruktionsergebnissen führt. Eine Messfeldvergrößerung durch eine geometrische Anpassung der Spulen ist somit möglich, jedoch im Ausmaß stark limitiert.

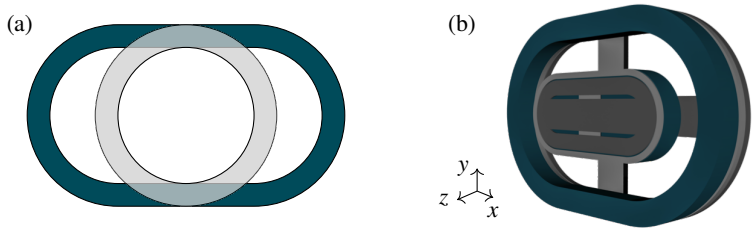



Abb. 5: (a) Illustration der geometrischen Veränderung einer kreisrunden Spule durch Änderung des Hauptachsenverhältnisses sowie (b) einer asymmetrischen Spulenanordnung auf Basis der geometrisch angepassten Spulen (jeweils aus [Ka17a]).

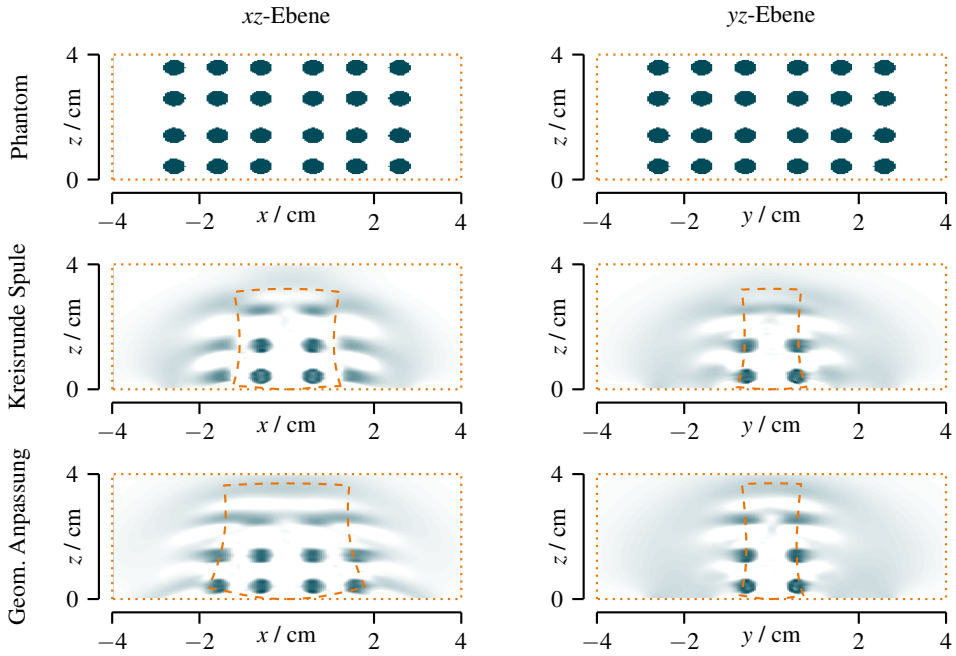



Abb. 6: Rekonstruktionsergebnisse in einer Schicht der *xz*- und der *yz*-Ebene unter Verwendung asymmetrischer Spulenanordnungen auf Basis kreisrunder und geometrisch angepasster Spulen (aus [Ka17a]).

Die Untersuchungen aus diesem Abschnitt haben somit gezeigt, dass eine Vergrößerung des Messfeldes in axialer Richtung bei gleichbleibender Auflösung des Systems nur bedingt durch eine geometrische Anpassung der Spulendimensionen erreichbar ist. Die untersuchte Änderung des Seitenverhältnisses kreisrunder Spulen bietet allerdings einen interessanten Ausgangspunkt für die Verbesserung von Spulenformen im Kontext von MPI und sollte somit in zukünftigen Entwicklungskonzepten berücksichtigt werden.

## 5.2 Elongationsbasierte Datenakquisition

Die Erweiterung des Messfeldes ist ein Bestreben, das vor allem im Hinblick auf potenzielle Applikationen eine wichtige Rolle in der aktuellen Forschung von MPI einnimmt. Der aktuell vielversprechendste Ansatz zur Erweiterung des Abtastbereiches ist die Verwendung von spezifischen Magnetfeldkonfigurationen. Eine technische Limitierung dieses Ansatzes besteht jedoch darin, dass eine Erzeugung solcher Felder eine axiale Beschränkung durch das MPI-System erfährt.

Alternativ hierzu konnte im Rahmen dieser Arbeit ein Verfahren vorgestellt werden, das über eine kontinuierliche Objektbewegung eine axiale Elongation der Abtasttrajektorie

(siehe Abb. 7 (a)) mit sich bringt und somit eine theoretisch axial unbeschränkte Vergrößerung des Messfeldes erlaubt. Aufgrund der Ähnlichkeit des vorgeschlagenen Akquisitionsschemas zu vergleichbaren Ansätzen in der CT-Bildgebung wurde sich zunächst mit einigen CT-Grundlagen sowie mit potenziell interessanten Aspekten für eine Übertragung auf MPI befasst. Anschließend wurde das generelle Prinzip der Trajektorienelongation in MPI beschrieben. Um eine Datenakquisition zu gewährleisten, die möglichst keinen Informationsverlust mit sich bringt, wurden eine physikalisch motivierte Elongationslänge sowie geeignete Berechnungsansätze formuliert. Aufbauend auf diesen Überlegungen wurden Untersuchungen durchgeführt, die sich mit Bildeinflüssen und insbesondere mit Artefakten, wie beispielsweise Verschmierungen im rekonstruierten Bild, durch elongierte Trajektorien befassen. Neben einer Bestätigung der physikalischen Annahmen hinsichtlich der Elongationslänge, konnte beispielsweise anhand einer Hauptkomponentenanalyse gezeigt werden, dass etwaig entstehende Verschmierungen stets richtungsabhägig sind und eine direkte Korrelation zur Elongation der Trajektorie aufweisen. Abschließend wurde ein Rekonstruktionsansatz basierend auf einer axialen Interpolation akquirierter Spannungssignale vorgeschlagen, der eine beliebige Schichtpositionierung auch nach der eigentlichen Aufnahme der Daten erlaubt (siehe Abb. 7 (b)). Zum Abschluss des Kapitels wurde sich ausführlich mit Analogien des neu eingeführten Bildgebungsschemas in MPI zu entsprechende Techniken der CT-Bildgebung befasst und diese diskutiert.

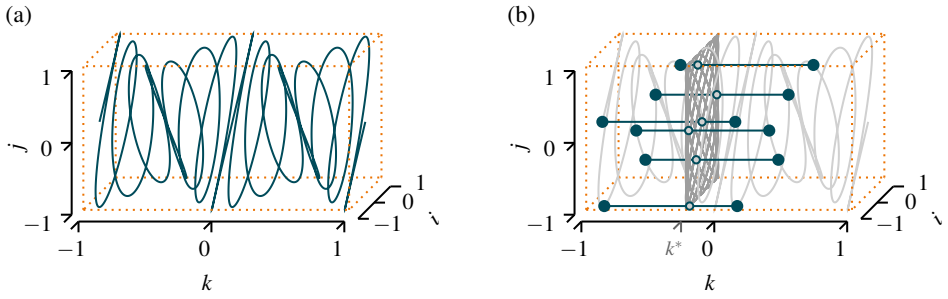



Abb. 7: (a) Darstellung einer axialen elongierten Lissajous-Trajektorie und (b) Veranschaulichung des Rekonstruktionsansatz basierend auf einer axialen Interpolation akquirierter Spannungssignale in eine gewählte Schicht (jeweils aus [Ka17a]).

Insgesamt konnte gezeigt werden, dass die Verwendung einer elongierten Trajektorie in MPI ein immenses Potenzial hinsichtlich der Bildgebung großer Messvolumen mit sich bringt. Eine exemplarisches Rekonstruktionsbeispiel unter Einhaltung der aufgestellten physikalisch Annahmen, das dieses Ergebnis unterstreicht, kann Abb. 8 entnommen werden.

## 6 Zusammenfassung und Ausblick

Durch die in der Promotionsarbeit von Christian Kaethner vorgeschlagenen Ansätze und Strategien konnten sowohl neue Erkenntnisse über die Nutzung und Erweiterung des Messfeldes in MPI gewonnen werden als auch neue Fragestellungen aufgeworfen werden, durch die sich wiederum vielfältige Ansatzpunkte für zukünftige Forschungsaktivitäten ergeben.

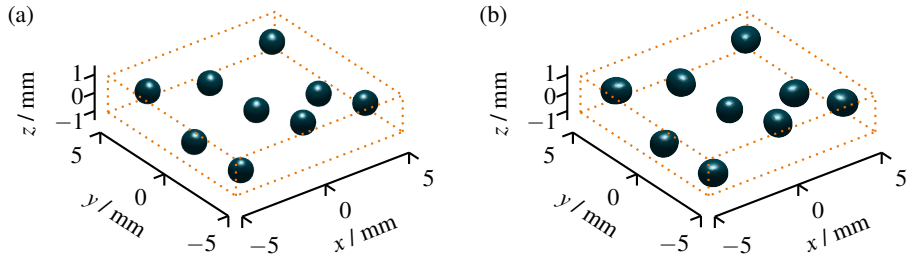



Abb. 8: (a) Phantom bestehend aus verschieden positionierten Kugeln sowie (b) beispielhaftes Rekonstruktionergebnis basierend auf einer Datenakquisition mit einer elongierten Lissajous-Trajektorie unter Einhaltung der physikalischen Annahmen (aus [Ka17a]).

Themenübergreifend über die vorgestellten Ansätze kann sich für nachfolgende Arbeiten beispielweise überlegt werden, wie die Konzepte der effizienten Gradientenfeldgenerierung mit denen einer elongierten Trajektorie vereinbar sind und ob die jeweilig verwendeten Magnetfeldkonfigurationen sich gegenseitig beeinflussen. Für die messfeldvergrößernde Bildgebgung mittels elongierter Trajektorien wäre es zudem interessant zu betrachten, wie sich eine nicht-äquidistante Akquisition einer MPI-Systemmatrix auf ein solches Akquisitionsschema übertragen lassen könnte. Die Bildeinflüsse könnten hierbei entsprechend unter Zuhilfenahme der vorgestellten Charakterisierungsmethoden auf Basis von Lissajous-Trajektorien vorgenommen werden.

## Literaturverzeichnis

[Er15a] Erb, W.; Kaethner, C.; Ahlborg, M.; Buzug, T. M.: Bivariate Lagrange interpolation at the node points of non-degenerate Lissajous curves. Numerische Mathematik, 133(4):685–705, 2015.

[Er15b] Erb, W.; Kaethner, C.; Denker, P.; Ahlborg, M.: A survey on bivariate Lagrange interpolation on Lissajous nodes. Dolomites Research Notes on Approximation, 8:23–36, 2015.

[Gr13] Grüttner, M.; Knopp, T.; Franke, J.; Heidenreich, M.; Rahmer, J.; Halkola, A.; Kaethner, C.; Borgert, J.; Buzug, T. M.: On the Formulation of the Image Reconstruction Problem in Magnetic Particle Imaging. Biomedizinische Technik/Biomedical Engineering, 58(6):583–591, 2013.

[GW05] Gleich, B.; Weizenecker, J.: Tomographic imaging using the nonlinear response of magnetic particles. Nature, 435(7046):1214–1217, 2005.

[Ka14] Kaethner, C.; Ahlborg, M.; Knopp, T.; Sattel, T. F.; Buzug, T. M.: Efficient gradient field generation providing a multi-dimensional arbitrary shifted field-free point for magnetic particle imaging. Journal of Applied Physics, 115(4):044910, 2014.

[Ka15a] Kaethner, C.; Ahlborg, M.; Bringout, G.; Weber, M.; Buzug, T. M.: Axially Elongated Field-Free Point Data Acquisition in Magnetic Particle Imaging. IEEE Transactions on Medical Imaging, 34(2):381–387, 2015.

[Ka15b] Kaethner, C.; Ahlborg, M.; Gräfe, K.; Bringout, G.; Sattel, T. F.; T. M. Buzug: Asymmteric Scanner Design for Interventional Scenarios in Magnetic Particle Imaging. IEEE Transactions on Magnetics, 51(2):6501904, 2015.

[Ka16] Kaethner, C.; Erb, W.; Ahlborg, M.; Szwargulski, P.; Knopp, T.; Buzug, T. M.: Non-Equispaced System Matrix Acquisition for Magnetic Particle Imaging based on Lissajous Node Points. IEEE Transactions on Medical Imaging, 35(11):2476–2485, 2016.

[Ka17a] Kaethner, C.: Strategien zur effzizienten Nutzung und Erweiterung des Messfeldes in Magnetic Particle Imaging. Dissertation, Universität zu Lübeck, 2017.

[Ka17b] Kaethner, C.; Cordes, A.; Hänsch, A.; Buzug, T. M.: Artifact Analysis for Axially Elongated Lissajous Trajectories in Magnetic Particle Imaging. International Journal on Magnetic Particle Imaging, 3(1):1703001, 2017.

[Kn09] Knopp, T.; Biederer, S.; Sattel, T.; Weizenecker, J.; Gleich, B.; Borgert, J.; Buzug, T. M.: Trajectory Analysis for Magnetic Particle Imaging. Physics in Medicine and Biology, 54(2):385–397, 2009.

[KSB12] Knopp, T.; Sattel, T. F.; Buzug, T. M.: Efficient Magnetic Gradient Field Generation With Arbitrary Axial Displacement for Magnetic Particle Imaging. IEEE Magnetics Letters, 3:6500104, 2012.

[MKB16] Mrongowius, J.; Kaethner, C.; Buzug, T. M.: Studies on the Improvement of Efficient Selection and Focus Field Coil Configurations. International Journal on Magnetic Particle Imaging, 2(1):1606001, 2016.

[Pa15] Panagiotopoulos, N.; Duschka, R.; Ahlborg, M.; Bringout, G.; Debbeler, C.; Graeser, M.; Kaethner, C.; Lüdtke-Buzug, K.; Medimagh, H.; Stelzner, J.; Buzug, T. M.; Barkhausen, J.; Vogt, F. M.; Haegele, J.: Magnetic Particle Imaging - Current Developments and Future Directions. International Journal of Nanomedicine, 10:3097–3114, 2015.

[Sa09] Sattel, T. F.; Knopp, T.; Biederer, S.; Gleich, B.; Weizenecker, J.; Borgert, J.; Buzug, T. M.: Single-Sided Device for Magnetic Particle Imaging. Journal of Physics D: Applied Physics, 42(2), 2009.

[We09] Weizenecker, J.; Gleich, B.; Rahmer, J.; Dahnke, H.; Borgert, J.: Three-dimensional real-time in vivo magnetic particle imaging. Physics in Medicine and Biology, 54(5):L1–L10, 2009.

[Zh15] Zheng, B.; Vazin, T.; Goodwill, P. W.; Conway, A.; Verma, A.; Saritas, E. U.; Schaffer, D.; Conolly, S. M.: Magnetic Particle Imaging tracks the long-term fate of in vivo neural cell implants with high image contrast. Scientific Reports, 5:14055, 2015.

**Christian Kaethner** wurde 1987 in Hannover geboren. Sowohl sein Bachelor- als auch sein Masterstudium absolvierte er an der Universität zu Lübeck im Bereich Medizinische Ingenieurwissenschaft und fokussierte sich hierbei auf verschiedene Aspekte der medizinischen Bildgebung und Bildverarbeitung. Daran anknüpfend promovierte er am dortigen Institut für Medizintechnik bei Prof. Dr. Thorsten M. Buzug im Bereich Magnetic Particle Imaging und schloss seine Promotion mit dem Prädikat *summa cum laude* ab.

# Ein Abstraktionsmodell für oberflächenbasierte gegenständliche Benutzerschnittstellen

Martin Kaltenbrunner[1]

**Abstract:** Die kumulative Dissertation mit dem englischen Originaltitel *An Abstraction Framework for Tangible Interactive Surfaces* diskutiert - am Beispiel vier aufeinander folgender Publikationen - die einzelnen Schichten eines gegenständlichen Interaktionsmodells, das im Zusammenhang mit einem elektronischen Musikinstrument mit gegenständlicher Benutzerschnittstelle entwickelt wurde. Basierend auf den Erfahrungen die während der Gestaltung und Implementierung dieser konkreten musikalischen Anwendung gesammelt wurden, konzentriert sich diese Forschungsarbeit hauptsächlich auf die Definition eines generell einsetzbaren Abstraktionsmodells für die digitale Repräsentation physischer Interfacekomponenten welche üblicherweise im Kontext interaktiver Oberflächen verwendet werden. Gemeinsam mit einer detaillierten Beschreibung des zugrundeliegenden Abstraktionsmodells, behandelt diese Dissertation auch dessen konkrete Implementierung in Form einer detaillierten Protokollsyntax, die das verbindende Element einer verteilten Architektur für die Realisierung von oberflächenbasierten gegenständlichen Benutzerschnittstellen darstellt. Die eigentliche Implementierung des vorgestellten Abstraktionsmodells als konkretes Toolkit besteht aus dem *TUIO Protokoll* und der damit verbundenen Computervision Anwendung für Objekt- und Fingergestentracking *reacTIVision*, gemeinsam mit deren primärer Anwendung in der Realisierung des *Reactable* Synthesizers. Die Dissertation schliesst mit einer Evaluierung und Erweiterung des ursprünglichen TUIO Modells, mit der Präsentation von TUIO2 - einem Abstraktionsmodell der nächsten Generation, das für eine weitergehendes Genre von gegenständlichen Interaktionsplattformen und die dazugehörigen Anwendungsszenarien entworfen wurde.

## Forschungsthese

Diese Dissertation dokumentiert den Forschungsprozess und die damit verbundene Design- und Entwicklungsarbeit mit dem Ziel folgende These zu etablieren:

> *Die Einführung einer Abstraktionsschicht in Form einer semantischen Klassifizierung und der digitalen Darstellung physischer Interfacekomponenten einer gegenständlichen Interaktionsumgebung, unterstützt die sensor- und anwendungsunabhängige Gestaltung von oberflächenbasierten gegenständlichen Benutzerschnittstellen, durch die Bereitstellung einer allgemeinen Infrastruktur für die Integration der zugrundeliegenden Interfacetechnologien mit übergeordneten anwendungsspezifischen Modellen.*

[1] Fakultät Medien, Bauhaus-Universität Weimar, martin.kaltenbrunner@ufg.at

## Forschungskontext

Das Paradigma gegenständlicher Interaktion (Tangible User Interfaces) repräsentiert heute ein relativ gut entwickeltes Forschungsfeld im Bereich der Mensch-Maschine Interaktion. Zahlreiche Anwendungsmodelle wurden etabliert um die fundamentalen Designprinzipien zu beschreiben, die aus der Forschungs- und Entwicklungspraxis der vergangenen Jahrzehnte hervorgegangen sind. Die meisten dieser Modelle konzentrieren sich jedoch auf die Darstellung allgemeiner Gestaltungsmuster oder semantischer Klassifikationen der Funktion, dem Inhalt oder der Beziehung physischer Designelemente auf der Anwendungsebene und aus der Perspektive der Benutzer.

Da gegenständliche Benutzerschnittstellen per Definition die physische Welt mit digitaler Information koppeln, vereinen die meisten aktuellen Konzepte die Beschreibung von physischen Interfacekomponenten mit ihrer jeweiligen Rolle auf der Anwendungsebene. Obwohl diese Modelle aus einer Designperspektive eine solide theoretische Grundlage für die Gestaltung gegenständlicher Interfaces durch ihre konzeptionelle Verkörperung digitaler Information in der Form physische Artefakte bieten, maskieren diese jedoch oft den jeweiligen technischen Hintergrund, der für die Realisierung derartig komplexer Hardware-Plattformen notwendig ist.

Andererseits können Entwickler heute auch auf eine Vielzahl von frei verfügbaren Hard- und Software Werkzeugen zurückgreifen, welche für die Herstellung der vielfältigen Manifestationen von physischen Interface-Komponenten verwendet werden können. Dies inkludiert Computer-Vision Anwendungen oder Unterhaltungselektronik mit fortschrittlicher Sensortechnologie, welche die notwendigen Kontrolldaten für die Integration physischer Interfacekomponenten in die Interaktionsebene zur Verfügung stellen. Da diese Werkzeuge und Geräte oft nur generische Sensordaten liefern, erfordert dies aus der Entwicklerperspektive eine zusätzliche Übersetzung der Rohdaten.

Diese Dissertation etabliert ein Modell für die semantische Beschreibung der Arten, Zustände und Beziehungen physischer Interfacekomponenten. Dieses Abstraktionsmodell liefert eine generische Beschreibung deren Zustände und Attribute an die Anwendungsebene, welche auf der zugrundeliegenden Semantik interaktiver Oberflächen aufbaut. Das Modell dient daher als eine vereinende Zwischenschicht zur Integration der physischen Umgebung mit einer digitalen Anwendung, und bietet daher eine durchgehend konsistente Repräsentation aus einer Entwickler-, Designer und Benutzerperspektive.

## Forschungsbeiträge

Diese Forschung wurde in aufeinanderfolgenden Schritten durchgeführt, welche das Design, die Implementierung, eine Anwendung und die abschliessende Evaluierung des ursprünglichen Modells beinhalten, und letztendlich zur Definition des erweiterten Abstraktionsmodell führten, welches im abschliessenden Kapitel dieser Dissertation behandelt wird. Die folgenden Beiträge sind das Ergebnis dieser einzelnen Forschungsphasen:

- Die Definition eines Abstraktionsmodells für die semantische Beschreibung der physischen Interaktionsumgebung, die als Mediator zwischen den physischen Interfacekomponenten und der digitalen Anwendungsebene dient, und damit eine sensor- und anwendungsunabhängige Klassifizierung der physischen und gestenbasierten Elemente im Kontext interaktiver Oberflächen bereitstellt.
- Die Implementierung des oben genannten Abstraktionsmodells innerhalb der offenen TUIO Protokolldefinition und einer zugehörigen Programmierschnittstelle, welche eine dezidierte Syntax und die semantischen Deskriptoren zur Kommunikation der Eigenschaften gegenständlicher Interfacekomponenten an die Anwendungsebene bereitstellen.
- Eine Hard- und Software Referenzplattform dieses physischen Abstraktionsmodells in Form der Computer-Vision basierten Anwendung reacTIVision, welche auch die konkrete Implementierung der oben genannten Protokollinfrastruktur integriert.
- Das Interaktionsdesign des gegenständlichen Musikinstruments Reactable, welches die praktischen Implikationen und die konkrete Anwendung des Modells im Bereich musikalischer Interfaces demonstriert, indem es die direkte Manipulation von Klang in der Form physischer Objekte auf einer Tischoberfläche ermöglicht.
- Die Weiterentwicklung des urpsrünglichen Abstraktionsmodells auf der Basis einer Analyse weiterer gegenständlicher Interaktionsplattformen und dessen Erweiterung und Implementierung anhand des TUIO 2.0 Protokolls, um die erweiterten Eigenschaften gegenwärtiger tischbasierter Interaktionsmodelle zu reflektieren.

## Forschungsmethoden

Während der Entwicklung des präsentierten Modells, seiner Protokoll- und Referenzimplementierung habe ich eine konsequente *Open Science* Methodik angewendet, indem ich eine *Public Domain* Protokollspezifikation gemeinsam mit einem *Open Source* Toolkit zur Verügung gestellt habe, welches unter Anderem auch die Aktiväten einer kreativen *Open Design* Community unterstützt hat. In diesem Kontext bezieht sich der Begriff *Public Domain* auf die lizenzfreie Verwendung des TUIO Protokolls, während das *Open Source* Konzept den vollständigen Zugriff auf den Quellcode seiner Implementierung garantiert. *Open Design* erweitert diese Idee auf weitergehende Details zu den spezifischen Hardwaredesign-Aspekten, die in Zusammenhang mit einer konkreten Anwendung dieser Technologien stehen. Dies stellt in Kombination mit einer formalen Veröffentlichung aller relevanten wissenschaftlichen Ergebnisse und einer entsprechenden Dokumentation das elementare *Open Science* Werkzeug im Bereich der Mensch-Maschine-Interaktion dar.

Diese offene Methodologie wurde zusätzlich durch das modulare Design des dokumentierten Frameworks erleichtert, wodurch auch die Integration von zahlreichen Communitybeiträgen in meine eigene Forschungspraxis ermöglicht wurde. Dieser Open Science Ansatz generiert ausserdem vergleichbare Resultate, welche auf einer gemeinsamen Forschungs-Infrastruktur aufbauen, die eine Evaluierung und die weitere Verbesserung individueller

Forschungsaspekte innerhalb eines geteilten Ökosystems erst ermöglichen. Diese Forschungspraxis besteht daher nicht nur aus einer *peer-reviewed* Veröffentlichung von wissenschaftlichen Ergebnissen, sondern zusätzlich auch aus der *peer-improved* Veröffentlichung einer entsprechenden Open-Source Implementierung. Ich erachte dies als die notwendigen Elemente für die Herstellung einer *Open-Experiment* Situation, auch mit Untersützung durch *Crowd-Source* Beiträge aus der Community.

Eine konkrete Anwendung wurde durch einen *praxisorientierten Forschungsansatz* innerhalb eines musikalischen Anwendungsfeldes im Reactable evaluiert, einem tischbasierten modularen Synthesizer mit gegenständlicher Benutzerschnittstelle. Dies inkludierte auch die Anwendung *künstlerischer Forschungsmethoden* durch die Integration von Komponisten und Performern in den Designprozess. Die Erfahrungen aus meiner eigenen Performancepraxis in Kombination mit dem Feedback aus den verschiedenen künstlerischen Kooperationen wurden in einem *Iterativen Designprozess* in die weitere Verbesserung der technischen Grundlagen und des Interaktionskonzeptes dieses Musikinstrumentes integriert.

Das finale Abstraktionsmodell dieser Dissertation beruht auf einer umfassenden Analyse und Klassifikation zahlreicher interaktiver Oberflächen und deren grundlegenden physischen Komponenten, und wurde auch durch eine exemplarische Kodierung dieser Interaktionsplattformen im Rahmen dieses Modells evaluiert.

## Ausgewählte Publikationen

Diese Dissertation ist in drei Abschnitte gegliedert. Der erste Abschnitt bietet eine allgemeine Einführung in den Forschungsbereich der gegenständlichen Interaktion und definiert die spezifischen Eigenschaften von interaktiven Oberflächen, welche den konkreten Anwendungsbereich meiner Forschung darstellen.

Die folgenden vier Artikel, welche ich als Erstautor in den Jahren von 2005 bis 2009 veröffentlicht habe, bilden den zentralen Anschnitt dieser kumulativen Dissertation. Laut Google Scholar hat diese Forschung seit ihrer ursprünglichen Veröffentlichung mittlerweile mehr als 1100 Zitate akkumuliert.[3]

[1] Martin Kaltenbrunner, Till Bovermann, Ross Bencina und Enrico Costanza. “TUIO - A Protocol for Table Based Tangible User Interfaces”. In: *Proceedings of the 6th International Workshop on Gesture in Human-Computer Interaction and Simulation (GW 2005)*. Vannes, France, 2005.

[2] Martin Kaltenbrunner, Sergi Jordà, Günter Geiger und Marcos Alonso. “The reacTable: A Collaborative Musical Instrument”. In: *Proceedings of the Workshop on Tangible Interaction in Collaborative Environments (TICE)*. Manchester, U.K., 2006.

[3] `http://scholar.google.com/citations?user=G7rN7JUAAAAJ`

[3] Martin Kaltenbrunner und Ross Bencina. “reacTIVision: A Computer-Vision Framework for Table-Based Tangible Interaction”. In: *Proceedings of the first international conference on Tangible and Embedded Interaction (TEI07)*. Baton Rouge, Louisiana, 2007.

[4] Martin Kaltenbrunner. “reacTIVision and TUIO: A Tangible Tabletop Toolkit”. In: *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces (ITS2009)*. Banff, Canada, 2009.

Diese Kernpublikationen stellen die Modelldefinition in der Form des TUIO Protokolls, seine Implementierung innerhalb des reacTIVision Frameworks, eine Anwendung im Rahmen des Musikinstruments Reactable und eine abschliessende Evaluierung des Abstraktionsmodells für gegenständliche interaktive Oberflächen in Form der weitergehenden Evolution des reacTIVision und TUIO Frameworks dar.

Jedem Publikationskapitel ist eine einleitende Klarstellung meiner eigenen Beträge im jeweiligen Forschungskontext vorausgestellt, und wird mit einer nachträglichen Analyse der ursprünglichen Publikation und weiterführenden Anmerkungen im Gesamtkontext abgeschlossen. Die Kommentare zur letzten Publikation in diesem Abschnitt beinhalten eine tiefergehende Analyse der Stärken und Schwächen des Modells im Zusammenhang aller vier Publikationen und dokumentiert auch zahlreiche Verbesserungen des reacTIVision und TUIO Frameworks seit ihrer originalen Veröffentlichung.

Diese Arbeit schliesst mit einem dritten Abschnitt, welcher auf den Forschungsergebnissen dieser Kernpublikationen aufbauend ein erweitertes Abstraktionsmodell und seine Implementierung in Form des TUIO 2.0 Protokolls definiert. Der Abschnitt enthält neben der vollständigen Spezifikation dieser neuen Protokollgeneration auch eine Reihe von Beispielkodierungen, welche dessen Potential anhand einiger gegenständlicher Interaktionsplattformen, Anwendungen und Geräte illustriert.

## Das TUIO Protokoll

Am Beginn dieser Dissertation steht die Veröffentlichung der ursprünglichen Spezifikation des TUIO 1.0 Protokolls, welche die Beschreibung und Übertragung einfacher Multipointer-Positionsdaten als generische *Cursor* Komponenten ermöglicht. Ausserdem beschreibt das Protokoll auch abstrakte Gegenstände als *Object* Komponenten, welche zusätzlich zu ihrer Oberflächenposition und ihrem Rotationswinkel auch durch eine eindeutige Identifikationsnummer unterschieden werden können. Mit diesen beiden Kernkomponenten lassen sich sowohl Multitouch-Interaktion als auch die Manipulation markierter Objekte im Kontext einer interaktiven Oberfläche realisieren. Mit einer späteren Publikation wurde dieses Modell noch um eine dritte *Blob* Komponente ergänzt, um auch die Beschreibung unmarkierter Objekte ihrer vereinfachten Geometrie erlaubt. Diese drei Komponenten stellen die Kernelemente der finalen TUIO 1.1 Spezifikation dar, welche trotz ihres einfachen Designs die Realisierung relativ komplexer Anwendungen für interaktive Oberflächen ermöglichte.

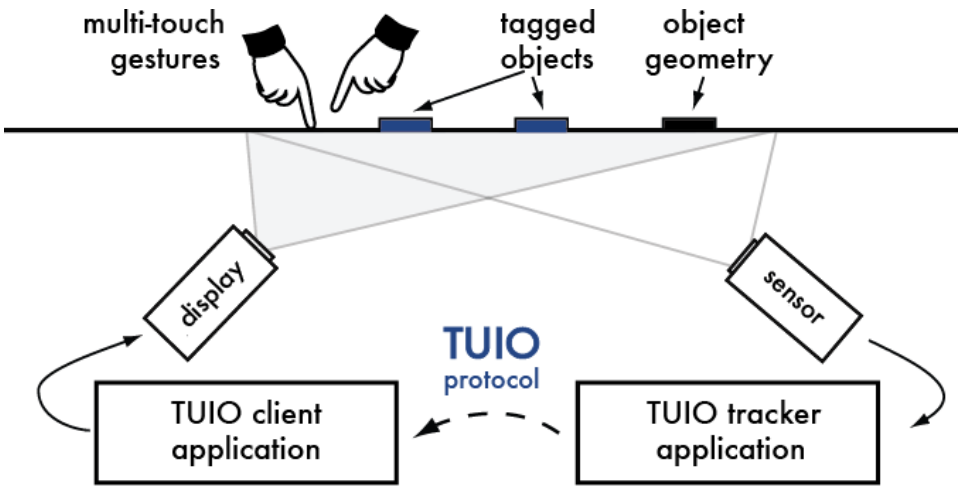



Abb. 1: Die verteilte TUIO 1.1 Architektur

Das TUIO Protokoll baut als verteilte Architektur auf dem *Open Sound Control* Format auf, und erlaubt daher eine getrennte Entwicklung der Hardware- und Anwendungsebene. Gemeinsam mit der frei verfügbaren Protokollspezifikation, wurden auch eine Reihe quelloffener Bibliotheken für die Integration des TUIO Protokolls in zahlreichen Programmiersprachen und Entwicklungsumgebungen zur Verfügung gestellt. Dieser Umstand führte letztendlich auch zu einer weiteren Verbreitung des Protokolls, und seiner Implementierung in zahlreichen Soft- und Hardwareprojektion für die Realisierung von interaktiven Oberflächen. Auch wenn das Protokoll selbst auch explizit Objektinteraktion ermöglichte, so konzentrierte sich ein Großteil der Entwicklungen vor allem auf Multitouch-Interaktion, da diese zu jenem Zeitpunkt von Standardsystemen noch nicht ausreichend unterstützt wurden. Auch wenn Das TUIO Protokoll in diesem konkreten Anwendungsbereich heute weitgehend obsolet geworden ist, so stellt seine zusätzliche Möglichkeit der Objektbeschreibung nach wie vor ein wichtiges Alleinstellungsmerkmal dar.

## Das reacTIVision Framework

Die zweite Publikation in der vorliegenden Dissertation beschreibt mit dem reacTIVision Framework eine Implementierung des TUIO Modells in der Form einer Computer-Vision Anwendung. reacTIVision wurde im Kern für das Tracking spezieller Markersymbole entwickelt, welche auch auf beliebige Gegenständen angebracht werden können. Dies erlaubt in der Folge die eindeutige Identifikation und exakte Lokalisierung dieser Objekte, sowie die Erfassung ihres Rotationswinkels. Zusütlich zum Symboltracking wurde reacTIVision auch um eine einfache Methode erweitert, die auch das Tracking von Fingerkuppen in Kontakt mit einer halbdurchlüssigen Oberäche erlaubt. In einem späteren Entwicklungsschritt wurde die Anwendung letzendlich auch noch um die Analyse generischer Objektgeometrien erweitert, womit reacTIVision heute eine vollständige Referenzimplementierung des TUIO 1.1. Abstraktionsmodells darstellt.

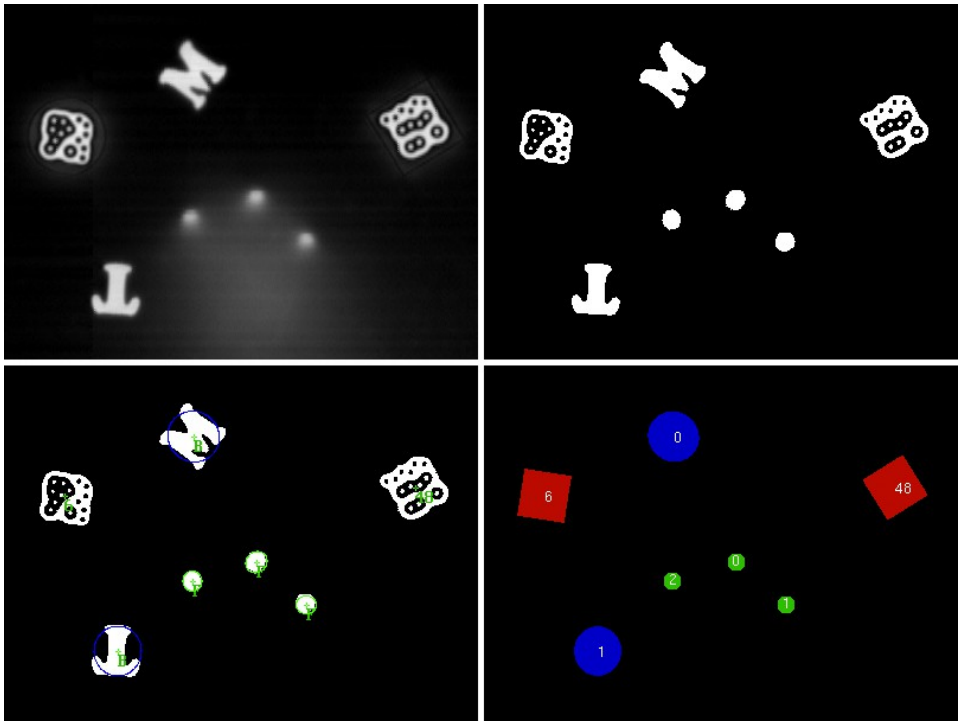



Abb. 2: Implementierung der TUIO 1.1 Abstraktion in reacTIVision

reacTIVision zeichnet sich aufgrund seines topologischen Markerdesigns durch eine relativ hohe Robustheit und Effizienz aus. Die Bildverarbeitung wurde dahingehend optimiert, dass diese lediglich aus einem Tresholding und Segmentierungsschritt besteht. Die Objekterkennung und die Blobanalyse finden bereits in dafür optimierten Datenstrukturen statt. Ein weiteres Unterscheidungsmerkmal der Software ist die spezielle Ästhetik der *Amoeba* Symbole, welche auf einem genetischen Algorithmus für das Rendering seiner optimierter Formen beruht. Die Software ist ebenfalls gemeinsam mit seinem Quellcode für mehrere Standardplattformen verfügbar und wurde in der Folge für die Umsetzung zahlreicher künstlerischer und wissenschaftlicher Projekte genutzt.

## Der Reactable Synthesizer

Der Reactable ist ein elektronisches Musikinstrument in der Form eines runden Tisches, auf dessen Oberfläche einfache geometrische Objekte arrangiert werden können. Diese Objekte repräsentieren die Kernelemente eines modularer Synthesizers, in der Form verschiedener Klangerzeuger, Filter und Klangeffekte, sowie Controllerobjekte und Sequencer. Im Prinzip stellt der Reactable eine Arte gegenständliche Klangprogrammiersprache dar, welche vor allem durch die Manipulation physischer Objekte und deren Beziehungen zueinander gestaltet werden kann. Im Rahmen der Entwicklung dieses Musik-

instrumentes war es auch notwendig geworden, die dafür notwendigen Basistechnologien zu entwickeln. Dies führte letztendlich auch zur Spezifikation des TUIO Protokolls und seiner Implementierung im reacTIVision Framework, welche im Wesentlichen die für die Realisierung des Reactables notwendigen Interface-Komponenten zur Verfügung stellen, sowie die Herstellung einer kamerabasierten interaktiven Oberflüche ermöglichte.


Abb. 3: Der Reactable in seiner heutigen Form.

Die in der Dissertation behandelte Publikation konzentriert sich vor Allem auf den kollaborativen Charakter dieses Musikinstrumentes, welcher unter Anderem auch durch seine spezielle runde Tischform betont wird. Das erste öffentliche Reactable-Konzert im Jahr 2005 wurde außerdem gleichzeitig an zwei Orten aufgeführt, wobei die beiden Instrumente über das Internet miteinander verbunden waren, und über eine Distanz von mehreren hundert Kilometern eine kooperative Performance innerhalb einer gemeinsam genutzten interaktiven Oberfläche stattfinden konnte. Das besonders robuste Design des TUIO Protokolls garantierte dabei die stabile Übertragung der entfernten Interaktionen und damit eine konsistente visuelle und akustische Darstellung des Instruments an beiden Konzertorten. Das TUIO Protokoll leitet das Erscheinen und Entfernen von Komponenten aus den Veränderungen einer Liste aller aktiven Elemente ab.

## Forschungsergebnisse

Basierend auf den umfassenden Erfahrungen, die ich im Rahmen der Entwicklung dieses Modells sowie seiner Implementierung als Teil eines gemeinschaftlich entwickelten Soft- und Hardware Ökosystems gesammelt habe, habe ich eine erweiterte und neue Generation dieses Abstraktionsmodells gestaltet, welche aktuelle Forschungsergebnisse und

Technologien für gegenständliche interaktive Oberflächen berücksichtigt. Ich hoffe, dass dieses aktualisierte Modell die weitere Forschung nicht nur im Bereich der gestenbasierten Multitouch-Interaktion unterstützt, sondern vor Allem auch die weitere Entwicklung des Paradigmas gegenständlicher Interaktion in der physischen und begreifbaren Domäne fördert.

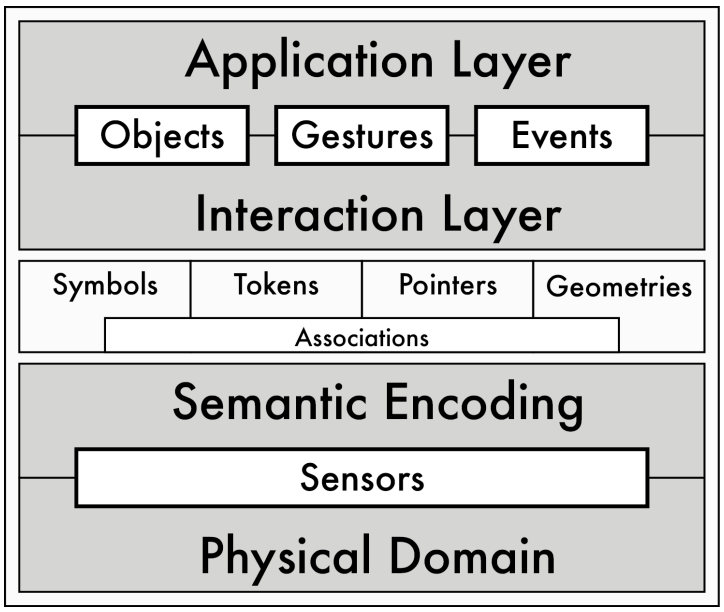



Abb. 4: Schichten des allgemeinen TUIO Abstraktionsmodells

Das **ursprüngliche Abstraktionsmodell** hat die drei grundlegenden Interfacekomponenten von **Tokens** (als Objekte), **Pointers** (als Cursors) und **Bounds** (als Blobs) definiert, welche die fundamentalen Elemente gegenständlicher interaktiver Oberflächen darstellen. Auf der Grundlage dieses Modells wurde das heute de-facto Standard **TUIO-Protokoll** definiert, welches die Zustände und Eigenschaften dieser Komponenten darstellt. Diese Protokollabstraktion stellt eine semantische Beschreibung physischer Interfacekomponenten unabhängig von den eigentlichen Hardwareeigenschaften und Sensortechnologien dar, und erlaubt damit die Entwicklung plattformunabhängiger Anwendungen.

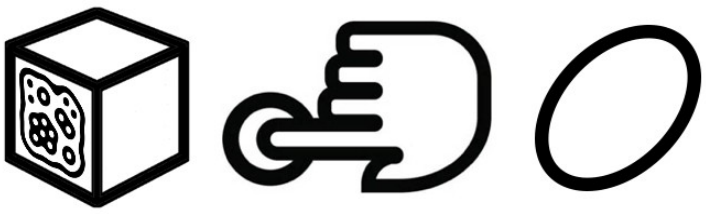


Abb. 5: TUIO 1.1 Komponenten: Objekte, Cursor & Blobs

Dafür habe ich auch eine umfassende Sammlung von Softwarewerkzeugen bereitgestellt, welche nicht nur das TUIO Protokoll implementieren sondern dessen Kernfunktionalität

auch in Form eines Computer-Vision Toolkits zur Realisierung tischbasierten Anwendungen illustrieren. Dieser **Open-Source Ansatz** führte auch zur Entwicklung zahlreicher weiterer auf diesem Protokoll aufbauender Werkzeuge und Anwendungen. Auch wenn die meisten dieser externen Lösungen die vielseitige Anwendbarkeit des generischen Abstraktionsmodells belegen, haben einige spezifische Anwendungsfälle auch diverse Einschränkungen dieser Vereinfachung aufgezeigt.

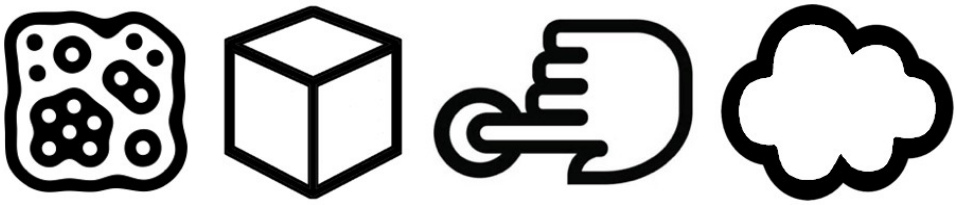


Abb. 6: TUIO2 Komponenten: Symbole, Token, Pointer & Geometrien

Die detaillierte Analyse von zahlreichen Communitybeiträgen und weiteren gegenständlichen Interaktionsplattformen, haben mich daher zur Definition eines **erweiterten Abstraktionsmodells** motiviert. Während **Tokens**, **Pointers** & **Bounds** mit einem erweiterten Set von Attributen nach wie vor die Kernkomponenten dieses neuen Abstraktionsmodells darstellen, habe ich eine zusätzliche generische **Symbol** Komponente eingeführt, sowie auch eine Reihe von **Geometrien**, welche die physische Objekterscheinung detailliert beschreiben. Das erweiterte Modell erlaubt ausserdem die Einbindung von **Control** Elementen, sowie die Beschreibung physischer oder logischer **Assoziationen** und den Austausch von **Signalen** zwischen Komponenten.

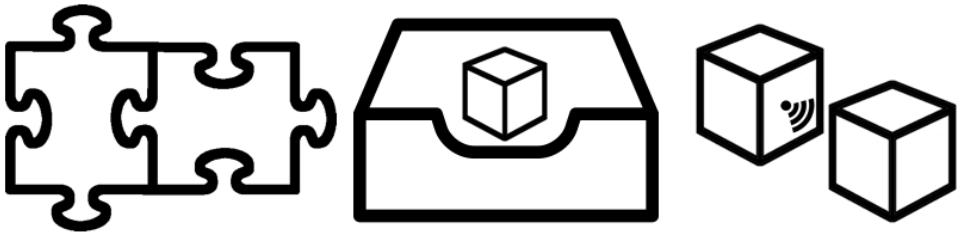


Abb. 7: TUIO2 Beziehungen: Physischer Link, Container & Signal

Es bleibt zu hoffen, dass das TUIO 2.0 Protokoll und die damit verbundenen Anwendungen, Bibliotheken und Werkzeuge eine ausreichend attraktive Funktionserweiterung darstellen um auch die weitere Community-Unterstützung dieser neuen Protokollgeneration zu gewährleisten, in ähnlicher Form wie die vielfältigen Beiträge zum bestehenden TUIO 1.1 Ökosystem.

Steffen Hölldobler et. al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2018 121

# Exakter Entwurf digitaler mikrofluidischer Biochips[1]

Oliver Keszöcze[2]

**Abstract:** Eine große Zahl von medizinischen Laboruntersuchungen oder biologischen Experimenten wird heutzutage mit großem Aufwand von spezialisierten Fachkräften und unter Zuhilfenahme von hochentwickelten, teuren Apparaturen durchgeführt. Dies führt zu hohen Kosten und einem geringen Probendurchsatz. Digitale mikrofluidische Biochips bieten hier eine vielversprechende Alternative. Mit ihrer Hilfe ist es möglich, diese Labortätigkeiten auf eine revolutionär andere Art und Weise, auf kleinstem Raum und ohne manuelle Interaktion zu realisieren („Labor-auf-einem-Chip"). Um ein Experiment auf einem solchen Biochip durchführen zu können, muss eine Reihe von Entwurfsproblemen gelöst werden (etwas das Platzieren von Modulen auf dem Biochip). Während diese Probleme bisher immer isoliert betrachtet und gelöst wurden, wird in der hier vorgestellten Arbeit erstmalig ein Ansatz präsentiert, welcher den gesamten Entwurf in einem Schritt betrachtet und umfassende Lösungen für alle Entwurfsprobleme gleichzeitig liefert. Damit werden nicht nur schwerwiegende Probleme vermieden, die bei der Aufteilung in Teilprobleme entstehen; die gefundenen Lösungen sind auch noch nachweislich am günstigsten.

## 1 Einführung

Viele biologische oder medizinische Experimente werden derzeit manuell von spezialisierten Fachkräften durchgeführt. Dies geschieht üblicherweise in einem Labor, welches mit umfangreicher und hochkomplexer Ausstattung versehen ist (Abb. 1(a) zeigt ein typisches Labor). Hierdurch wird der gesamte Prozess teuer und erreicht keinen hohen Durchsatz. Darüber hinaus sind insbesondere monotone und repetitive Vorgänge bei manueller Ausführung fehleranfällig.

Diese Problematik führte zur Entwicklung von (voll-)automatisierten Laborgeräten (siehe Abb. 1(b)). Diese Geräte erreichen einen hohen Automatisierungs- und Integrationsgrad, obwohl sie häufig die physischen Arbeitsvorgänge der Fachkräfte imitieren. Auch wenn sie die Laborarbeit signifikant erleichtern, sind solche Geräte immer noch groß und teuer.

Um die Größe von Laborgeräten weiter reduzieren zu können, wurde untersucht, wie sich Flüssigkeiten im Nano- bis Pikoliter-Bereich manipulieren lassen. Hieraus entstanden die mikrofluidischen Biochips (siehe Abb. 1(c)[3]), welche auch „Labor-auf-einem-Chip" (lab-on-a-chip) genannt werden. Sie sparen nicht nur Flüssigkeiten (welche teuer oder schwierig zu beschaffen sein können), sondern können auf Grund der geringeren Flüssigkeitsvolumina die gesamte Dauer des Experiments reduzieren.

[1] Englischer Titel der Dissertation: „Exact Design of Digital Microfluidic Biochips"

[2] Arbeitsgruppe Rechnerarchitektur (Leitung Prof. Dr. Rolf Drechsler), Universität Bremen, keszcoze@uni-bremen.de

[3] Der abgebildete Biochip entstand im Rahmen einer mitbetreuten Bachelorarbeit [Lü17]. Er ist aber auf Grund der inhaltlichen Ausrichtung der Arbeit auf den Entwurfsprozess nicht Teil der Dissertation.


(a) Labor (Größe: Raum).


(b) autom. Gerät (Größe: $m^3$).

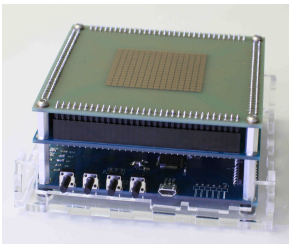

(c) Biochip (Größe: $cm^3$).

Abb. 1: Entwicklung der Gerätegröße.

Die Fähigkeiten von mikrofluidischen Geräten sind in der Literatur vielfach demonstriert worden. So lässt sich zum Beispiel die Polymerase-Kettenreaktion multiplex in Echtzeit durchführen [Li04]. Ein weiterer Bereich, in dem Biochips von großem Interesse sind, ist die Vorbereitung von Proben (siehe z.B. [Bh17]). Mit Hilfe von Biochips kann dieser mühsame Vorgang zu einem hohen Grade automatisiert werden. Wie in [Al17] gezeigt wurde, können Biochips die Zukunft für leicht zugängliche medizinische Versorgung sein. Ein mögliches Anwendungsszenario ist der Einsatz von Biochips zum Test auf Krankheiten in abseits gelegenen oder nur schwer erreichbaren Regionen.

Ein zentrales Problem bei dem Einsatz von Biochips ist die Frage, wie ein gegebenes Experiment konkret auf einem Biochip ausgeführt werden kann. Dieser Entwurfsprozess ist das Thema der Dissertation.

## 2 Zentrale Beiträge der Dissertation

Der wesentliche Beitrag der Dissertation besteht darin, zu erkennen, dass der bisherige Entwurfsprozess von Biochips zwangsweise Probleme mit sich bringt, und aufzuzeigen wie diese durch den, in der Arbeit entwickelten, „Ein-Schritt“-Ansatz vollständig gelöst werden können. Die vorgestellte Lösung [Ke14] ist exakt in dem Sinne, dass sie nachweislich die günstigsten (d.h. kürzesten) Lösungen erzeugt. Zu Referenzzwecken wurde auch eine heuristische Variante des „Ein-Schritt“-Ansatzes implementiert [Wi15].

Die Domäne Biochips nebst zugehörigen Entwurfsproblemen wird mittels eines, im Rahmen der Dissertation entwickelten, formalen Modells repräsentiert. Dieses Modell erlaubt es, theoretische Betrachtungen durchzuführen. Bislang wurde in der Literatur, im Wesentlichen wegen des Fehlens eines formalen Modells, die Komplexität der Problemstellung nur vermutet. In der Arbeit konnte gezeigt werden, dass der Entwurfsprozess für Biochips NP-schwer ist [Ke18a].

Des Weiteren ist das gewählte Modell allgemein genug formuliert, dass sich die vorgestellten Entwurfslösungen leicht auf andere Typen von Biochips übertragen lassen[Sc17b, Ke17, Ke18b].

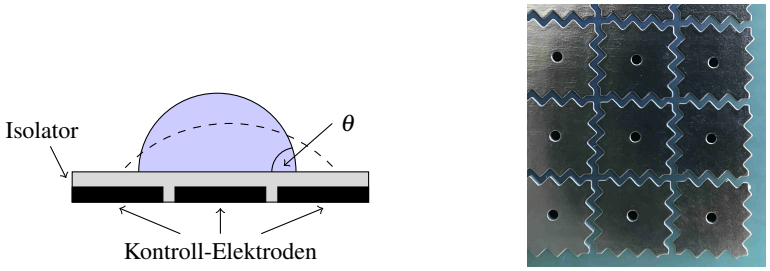
# 3 Digitale mikrofluidische Biochips

## 3.1 Technischer Hintergrund

Digitale mikrofluidische Biochips (digital microfluidic biochips, DMFBs) arbeiten mittels Elektrobenetzung (electrowetting-on-dielectric effect, siehe [PSF02]). Dieser Effekt ist in Abb. 2(a) illustriert. Ein Tropfen einer Flüssigkeit befindet sich auf einer hydrophoben Oberfläche, die den Tropfen elektrisch von Kontroll-Elektroden isoliert. Wird nun eine Spannung an eine der Elektroden angelegt, ändert sich der Kontaktwinkel zwischen Tropfen und Isolationsschicht wie in der Abbildung durch die gestrichelte Linie dargestellt wird. In der Abbildung ist das Resultat einer „angeschalteten“ rechten Elektrode visualisiert. Die Spannung würde dafür sorgen, dass sich der Tropfen über die rechte Elektrode bewegt.

Diese Elektroden, in der Literatur *Zellen* genannt, werden üblicherweise in quadratischer Form und in schachbrettartigem Layout auf einer Platine gefertigt. Abbildung 2(b) zeigt eine Aufnahme einer solchen Platine. Es ist zu beachten, dass, anders als in der schematischen Darstellung in Abb. 2(a), die Zellen gezackte Ränder haben. Dies ist nötig, da sich ein Tropfen zumindest partiell über einer Elektrode befinden muss, damit die Elektrobenetzung funktionieren kann. Durch ineinander greifende Zellen kann so die für einen Zellenwechsel nötige Größe des Tropfens klein gehalten werden.

Die Bewegungsdauer eines Tropfens wird abstrahiert. Die Bewegung eines Tropfens von einer Zelle zur nächsten ist ein Zeitschritt. Befinden sich mehrere Tropfen auf einem Biochip, können sich alle Tropfen innerhalb eines Zeitschrittes genau eine Zelle weit bewegen (oder auf ihrer aktuellen Position verharren). Die Bewegung eines Tropfens kann nur horizontal oder vertikal erfolgen. Das „digital“ in DMFB rührt daher, dass diskrete Volumina von Flüssigkeiten in Form von Tropfen bewegt werden.




(a) Elektrobenetzung. (b) Zellen auf einer Platine.

Abb. 2: **(a)** Illustration der Elektrobenetzung nach [PSF02]. Der Kontaktwinkel $\theta$ verändert sich mit angelegter Spannung. Der Tropfen verändert seine Form, wie durch die gestrichelte Linie angedeutet wird. Bild mit Änderungen übernommen aus [PSF02]. **(b)** Platine mit quadratischen Zellen mit gezackten Rändern.

### 3.2 Entwurf von Biochips

Der essentielle Schritt für die Nutzung von DMFBs wird Synthese oder Entwurf genannt. Das Ziel des Entwurfs ist es, ein biologisches oder medizinisches Experiment auf einem gegebenen Biochip zu realisieren. Ein Experiment wird als ein gerichteter Graph, Sequenzgraph genannt, modelliert, dessen Kanten die Abhängigkeiten zwischen den auszuführenden Operationen darstellen. Operationen sind hierbei z.B. das Vermischen zweier Flüssigkeiten oder das Erhitzen einer Probe. Es müssen zusätzlich das Layout des Biochips sowie weitere Kriterien, wie z.B. eine Begrenzung der Ausführungszeit, berücksichtigt werden. Mit diesen Eingaben und Einschränkungen wird ein konkreter Versuchsablauf erzeugt. Insgesamt müssen die folgenden Fragen adressiert werden:

- Welche Module werden genutzt, um eine Operation zu auszuführen? (*binding*)
- Wann (in welchem Zeitschritt) sollen die Operationen ausgeführt werden? (*scheduling*)
- Wo (auf welchen Zellen) sollen die Operationen ausgeführt werden? (*placement*)
- Welche Wege müssen die einzelnen Tropfen zurücklegen, um zu den Zellen ihrer jeweiligen Operationen zu gelangen? (*routing*)
- Welche Elektroden können gruppiert werden, um durch eine einfachere Kontroll-Logik gesteuert zu werden? (*pin assignment*)

Diese einzelnen Schritte werden in der Literatur in den Entwurf auf Architektur-Ebene (binding, scheduling) und den Entwurf auf physischer Ebene (placement, routing, pin assignment) unterteilt. Die erste Ebene beschäftigt sich abstrakt mit der Ausführung während die zweite Ebene konkrete Entitäten auf dem Biochip selbst behandelt. Der gesamte Entwurfsablauf ist in Abb. 3 dargestellt.

Die einzelnen Schritte sind ähnlich denen im klassischen Entwurf für Mikrochips. Allerdings gibt es einige Besonderheiten, die spezifisch für DMFBs sind.

Beim Entwurf kann man zwischen *statischen* und *dynamischen* Operationen unterscheiden. Zu den statischen Operationen gehören z.B. Detektoren, welche Eigenschaften der Flüssigkeiten analysieren. Sie sind fest in der Hardware des Biochips verbaut und können nicht mehr verändert werden. Eine dynamische Operation wie z.B. das Mischen zweier Flüssigkeitstropfen kann an einer beliebigen, freien Stelle auf dem Biochip ausgeführt werden. Außerdem wird die genutzte Fläche nach erfolgreicher Ausführung wieder freigegeben und steht nachfolgenden Operationen zur Verfügung. Im Gegensatz zum klassischen Routing auf Platinen, dürfen sich die Routen zweier Tropfen kreuzen, so lange sich die Tropfen zu keinem Zeitpunkt zu nahe kommen und sich nicht unbeabsichtigt vermischen.

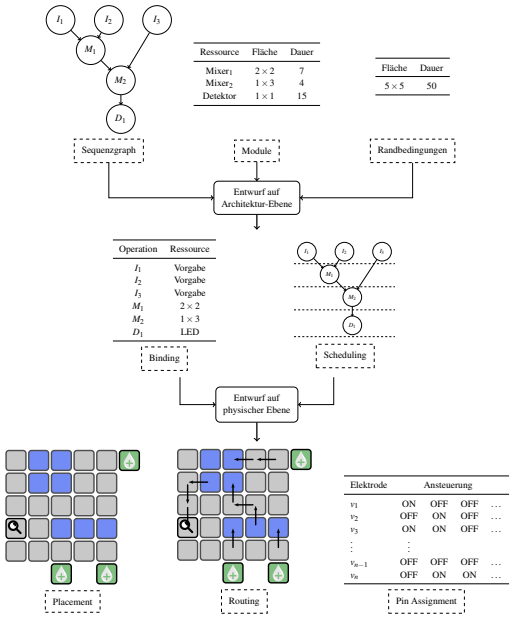



Abb. 3: Herkömmlicher Entwurfsablauf für DMFBs.

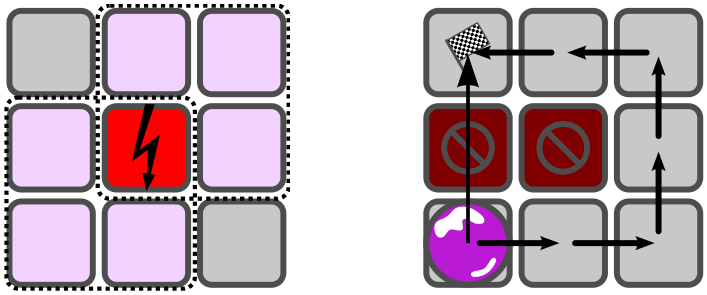
### 3.3 Problematik des sequentiellen Entwurfs

Bislang wurden die einzelnen Schritte binding, scheduling, placement, routing und pin assignment hauptsächlich voneinander isoliert in einem sequentiellen Entwurfsablauf gelöst (siehe hierzu exemplarisch [Ch13, SHC06, XC06]). Einige Arbeiten berücksichtigen zwar den direkten Vorgänger- oder Nachfolgerschritt, keine umfasst jedoch alle notwendigen Schritte auf einmal.

Dadurch, dass Teilschritte isoliert betrachtet werden, können zwei Probleme auftreten:

1. Sackgassen: Die Lösung eines Teilschrittes verhindert vollständig die Lösung eines darauf folgenden Schrittes
2. Suboptimale Lösungen: Eine Lösung eines Teilschrittes erzwingt eine unnötig teure Lösung eines darauf folgenden Schrittes (z.B. Lösungen mit vielen Zeitschritten)

Die erste Situation ist in Abb. 4(a) illustriert. Hier wurden Misch-Operationen (Mixer) platziert, die rechnerisch weniger Zellen benötigen, als der Biochip hat, die jedoch auf Grund ihrer Geometrie nicht auf das Gerät passen. Die zweite Situation ist in Abb. 4(b) dargestellt. Hier wurde ein Mixer so ungünstig platziert, dass ein Tropfen einen Umweg nehmen muss und somit die Anzahl der Zeitschritte unnötig hoch ist. Es gilt zu beachten, dass selbst dann unnötig hohe Kosten entstehen können, wenn die einzelnen Schritte mit sogenannten „exakten“ Verfahren gelöst werden, welche die niedrigsten Kosten für das jeweilige Problem garantieren.



(a) Platzierungsfehler (b) Unnötig teure Lösung

Abb. 4: **(a)** Es wurden zwei $2 \times 2$ Mixer ausgewählt, welche gleichzeitig ausgeführt werden sollen. Rechnerisch nutzen sie 8 von 9 zur Verfügung stehenden Zellen. Die konkreten Mixer-Geometrien lassen jedoch keine überlappungsfreie Platzierung zu. Das mit einem Blitz markierte Feld wird immer von beiden Mixern verwendet. **(b)** Ein im vorherigen Schritt platzierte Mixer blockiert die direkte Route des Tropfens zu seiner Zielzelle. Hierbei handelt es sich um eine vermeidbare suboptimale Lösung.

Eine lange Ausführungsdauer führt in der Praxis zu mehreren Problemen. Je nach konkreter Realisierung des Biochips besteht wegen der geringen Volumina der Tropfen die reale Gefahr, dass ein Tropfen verdunstet, ehe das Experiment erfolgreich abgeschlossen wurde. Außerdem können sich Eigenschaften der Flüssigkeiten ändern. Werden zum Beispiel Enzyme transportiert, können diese über die Zeit degenerieren. Dies macht es essentiell, dass im Entwurfsablauf Lösungen erzeugt werden, die eine möglichst geringe Laufzeit haben.

Sackgassen können dadurch behoben werden, dass man einen oder mehrere der vorherigen Schritte wiederholt und die fehlerhafte Lösung explizit ausschließt. Lösungen, die viele Zeitschritte benötigen, können nicht so einfach umgangen werden. Unter Umständen ist nicht einmal offensichtlich, dass eine Lösung mit weniger Zeitschritten existiert.

## 4 Exakter „Ein-Schritt“-Entwurf

Um die im vorherigen Abschnitt angesprochenen Probleme – Sackgassen und suboptimale Lösungen – anzugehen, wurde in der Dissertation erstmalig ein umfassender Ansatz gewählt. Die Idee ist, die Probleme an den Übergängen zwischen den Teilproblemen zu verhindern, in dem es keine Übergänge mehr gibt. Im „Ein-Schritt“-Entwurf werden alle Teilprobleme zu einem gemeinsamen Problem zusammengeführt und in einem Schritt gelöst.

Dieser Ansatz verhindert grundsätzlich Fehler zwischen einzelnen Schritten. In der Dissertation wurden zwei unterschiedliche Implementierungen dieses Ansatzes präsentiert. Neben einer heuristischen Methode wurde ein exakter Ansatz entwickelt, der die Minimalität der Lösung garantiert. Das Vorgehen hierbei ist, aus einem Entwurfsproblem eine Sequenz von Entscheidungsproblemen (satisfiability problem, SAT, siehe [Co71]) zu erzeugen. Hierbei wird das Entwurfsproblem in ein SAT-Problem überführt, dessen

Erfüllbarkeit äquivalent dazu ist, dass das ursprüngliche Problem eine Lösung besitzt. In dieser Sequenz wird der zu optimierende Parameter (meist die Anzahl der Zeitschritte) inkrementiert bis eine Lösung gefunden ist. Dies garantiert die Minimalität der Lösung. Die SAT-Probleme werden an spezielle SAT-Beweiser übergeben, welche effizient entscheiden können, ob eine Lösung existiert. Aus der Lösung des SAT-Problems wird die Lösung des ursprünglichen Entwurfsproblems extrahiert. Diese Grundidee ist in Abb. 5 visualisiert.

Ein Vorteil des Überführens in ein SAT-Problem ist, dass das Verfahren sehr dynamisch ist. Das betrachtete Entwurfsproblem lässt sich einfach erweitern und zusätzliche Randbedingungen können mit in das SAT-Problem kodiert werden. So lassen sich, z.B., unterschiedliche Zellgeometrien, wie dreieckige oder sechseckige Formen, betrachten (siehe [Sc17a]) oder der Einfluss von nur temporär blockierten Zellen auf den Entwurf untersuchen (siehe [KWD14]).

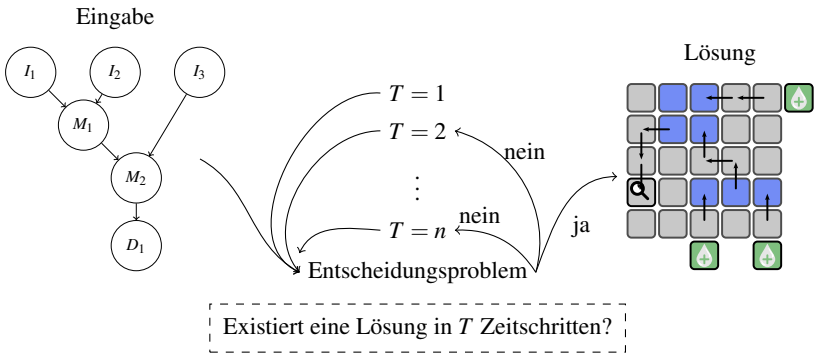



Abb. 5: Idee des exakten Ansatzes: Das Entwurfsproblem wird in eine Sequenz von Entscheidungsproblemen transformiert. Durch sukzessives Inkrementieren von $T$ wird garantiert eine Lösung mit der geringsten Anzahl an Zeitschritten gefunden.

### 4.1 Evaluation des exakten „Ein-Schritt“-Entwurfs

Die Effektivität des exakten „Ein-Schritt“-Ansatzes lässt sich gut mittels Experimenten aus der multiplex In-Vitro Diagnostik illustrieren. Hierbei werden $n$ Proben (samples) mit jeweils $m$ unterschiedlichen Indikatoren (reagents) vermischt und anschließend analysiert. Dieses Experiment lässt in hohem Maße Parallelität von Operationen zu. Entwurfsansätze sollten in der Lage sein, dies zu berücksichtigen, um Lösung mit möglichst wenigen Zeitschritten zu finden.

Experimente mit unterschiedlichen Kombinationen von Proben- und Indikatoranzahl (#$S$ bzw. #$R$) sowie unterschiedliche Biochip-Größen ($W \times H$) wurden sowohl exakt als auch heuristisch gelöst. Die Ergebnisse sind in Tab. 1 dargestellt. Die Spalte $T$ gibt die Anzahl der in den Lösungen verwendeten Zeitschritten an. Der heuristische Ansatz wurde 300 Mal pro Benchmark ausgeführt und die jeweils besten ($T^*$) und schlechtesten ($T^\dagger$) Lösungen aufgelistet. Zudem ist jeweils die Laufzeit des Ansatzes angegeben.

Für kleine Probleminstanzen erzeugt der heuristische Ansatz Lösungen, deren Anzahl an benötigten Zeitschritten nahe an denen des exakten Ansatzes liegt. Sobald jedoch die Problem- oder Biochip-Größe zunimmt, benötigen die besten heuristischen Lösungen ungefähr doppelt so viele Zeitschritte wie die des exakten Ansatzes. Die Laufzeit des exakten Ansatzes ist, wie beim SAT-Lösen zu erwarten, deutlich höher als die es heuristischen Ansatzes. Es konnte gezeigt werden, dass die beiden, in der Dissertation entwickelten, Ansätze geeignet sind, Entwurfsaufgaben mit hoher praktischer Relevanz zu lösen.

Tab. 1: Vergleich des exakten und des heuristischen „Ein-Schritt"-Verfahren anhand von multiplex In-Vitro Diagnostik Benchmarks.

| | | | Exakt | | Heuristisch | | |
|---|---|---|---|---|---|---|---|
| #S | #R | $W \times H$ | T | Dauer $(s)$ | $T^*$ | $T^\dagger$ | Dauer $(s)$ |
| 2 | 1 | $4 \times 6$ | 14 | 51.0 | 17 | 29 | 1.4 |
| 2 | 1 | $5 \times 5$ | 14 | 64.1 | 18 | 32 | 1.4 |
| 2 | 1 | $5 \times 6$ | 14 | 87.5 | 18 | 35 | 1.5 |
| 2 | 1 | $6 \times 6$ | 14 | 185.2 | 18 | 31 | 1.4 |
| 2 | 2 | $2 \times 5$ | 16 | 66.9 | 29 | 109 | 1.6 |
| 2 | 2 | $2 \times 6$ | 16 | 310.3 | 28 | 102 | 1.5 |
| 2 | 2 | $5 \times 5$ | 15 | 503.2 | 23 | 51 | 1.5 |
| 2 | 2 | $5 \times 6$ | 15 | 768.2 | 24 | 50 | 1.5 |
| 2 | 2 | $6 \times 6$ | 15 | 1262.0 | 24 | 87 | 1.5 |
| 2 | 3 | $3 \times 6$ | 17 | 3349.9 | 33 | 79 | 2.1 |
| 2 | 3 | $4 \times 6$ | 16 | 1122.9 | 32 | 56 | 1.6 |
| 2 | 3 | $5 \times 6$ | 16 | 1874.1 | 33 | 65 | 1.6 |
| 2 | 3 | $6 \times 6$ | 16 | 2147.6 | 33 | 76 | 1.7 |

### 4.2 Komplexität des Entwurfsproblems

In der Dissertation wurde die Komplexität von zwei Teilproblemen, dem Routing-Problem und dem Pin-Assignment-Problem, explizit nachgewiesen. Ihre Zugehörigkeit zur Komplexitätsklasse NP wurde vielfach in der Literatur vermutet (siehe [SHC06] und [XC06] für die jeweiligen Aussagen zum Routing-Problem und Pin-Assignment-Problem), jedoch nie bewiesen.

**Theorem 1.** *Das Routing-Problem und das Pin-Assignment-Problem für DMFBs sind* NP*-vollständig [Ke18a].*

Da Gesamtlösungen des Entwurfsproblems auch valide Lösungen für das Routing-Problem sowie das Pin-Assignment-Problem beinhalten, ist der „Ein-Schritt"-Ansatz mindestens so komplex wie Ansätze für diese beiden Teilprobleme. Damit ergibt sich das folgende Theorem.

**Theorem 2.** *Das DMFB Entwurfsproblem ist* NP*-schwer.*

Somit ist nachgewiesen, dass es im Allgemeinen nicht möglich ist, eine exakte Lösung schnell zu berechnen. Damit ist die Entscheidung, einen SAT-Beweiser zu verwenden, dem Entwurfsproblem angemessen.

## 5 Zusammenfassung

Biochips versprechen, die Durchführung von biologischen oder medizinischen Experimenten zu vereinfachen, zu beschleunigen und im Allgemeinen kostengünstiger zu machen. Darüber hinaus bieten sie große Chancen, in abgelegenen Gebieten die medizinische Versorgung zu verbessern. Dies ist möglich, da ein fertiger Biochip schon nach kurzer Einweisung auch von nicht speziell ausgebildeten Fachkräften zu bedienen ist. Bisherige Ansätze zum Entwurf von Biochips betrachteten das Problem als Abfolge voneinander unabhängiger Teilprobleme. Dies führt während des Entwurfs zu Problemen wie Sackgassen oder unnötig teuren Lösungen. Die Dissertation geht dieses Problem dadurch an, dass erstmalig alle Teilprobleme in einem „Ein-Schritt"-Verfahren gleichzeitig behandelt werden. Dadurch werden Probleme an den Übergängen zwischen Schritten vollständig verhindert. Darüber hinaus wird das Entwurfsproblem mit exakten Methoden, hier durch Verwendung von SAT-Beweisern, garantiert zu niedrigsten Kosten gelöst.

## Literatur

[Al17] Alistar, Mirela; Madsen, Jan; Ho, Tsung-Yi; Wille, Robert: When Embedded Systems meet Life Sciences: Microfluidic Biochips for Real-Time Healthcare. Bericht, 2017. Online abrufbar unter `https://www.researchgate.net/profile/Mirela_Alistar2`.

[Bh17] Bhattacharjee, Sukanta; Poddar, Sudip; Roy, Sudip; Huang, Juinn-Dar; Bhattacharya, Bhargab B.: Dilution and Mixing Algorithms for Flow-Based Microfluidic Biochips. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 36(4):614–627, 2017.

[Ch13] Chen, Ying-Han; Hsu, Chung-Lun; Tsai, Li-Chen; Huang, Tsung-Wei; Ho, Tsung-Yi: A Reliability-Oriented Placement Algorithm for Reconfigurable Digital Microfluidic Biochips Using 3-D Deferred Decision Making Technique. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 32(8):1151–1162, 2013.

[Co71] Cook, Stephen A.: The Complexity of Problem-Solving Procedures. In: ACM Symposium on Theory of Computing. ACM, S. 151–158, 1971.

[Ke14] Keszocze, Oliver; Wille, Robert; Ho, Tsung-Yi; Drechsler, Rolf: Exact One-pass Synthesis of Digital Microfluidic Biochips. In: Design Automation Conference. DAC, S. 142:1–142:6, 2014.

[Ke17] Keszocze, Oliver; Li, Zipeng; Grimmer, Andreas; Wille, Robert; Chakrabarty, Krishnendu; Drechsler, Rolf: Exact Routing for Micro-Electrode-Dot-Array Digital Microfluidic Biochips. In: Asia and South Pacific Design Automation Conference. ASP-DAC, S. 708–713, 2017.

[Ke18a] Keszocze, Oliver; Friedemann, Arved; Niemann, Philipp; Drechsler, Rolf: On the Complexity of Design Tasks for Digital Microfluidic Biochips. 2018. (under review).

[Ke18b] Keszocze, Oliver; Ibrahim, Mohamed; Wille, Robert; Chakrabarty, Krishnendu; Drechsler, Rolf: Exact Synthesis of Biomolecular Protocols for Multiple Sample Pathways on Digital Microfluidic Biochips. In: International Conference on VLSI Design. S. 121–126, 2018.

[KWD14] Keszocze, Oliver; Wille, Robert; Drechsler, Rolf: Exact Routing for Digital Microfluidic Biochips with Temporary Blockages. In: International Conference On Computer Aided Design. ICCAD, S. 405–410, 2014.

[Li04] Liu, Robin Hui; Yang, Jianing; Lenigk, Ralf; Bonanno, Justin; Grodzinski, Piotr: Self-Contained, Fully Integrated Biochip for Sample Preparation, Polymerase Chain Reaction Amplification, and DNA Microarray Detection. Analytical Chemistry, 76(7):1824–1831, 2004.

[Lü17] Lünert, Maximilian: StackADrop: A versatile Biochip. 2017.

[PSF02] Pollack, Michael G.; Shenderov, Alexander D.; Fair, Richard B.: Electrowetting-based actuation of droplets for integrated microfluidics. Lab on a Chip, 2:96–101, 2002.

[Sc17a] Schneider, Leonard; Keszocze, Oliver; Stoppe, Jannis; Drechsler, Rolf: Der Einfluss von Zellformen auf das Routing von Digital Microfluidic Biochips. In: Methoden und Beschreibungssprachen zur Modellierung und Verifikation von Schaltungen und Systemen. MBMV, S. 75–77, 2017.

[Sc17b] Schneider, Leonard; Keszocze, Oliver; Stoppe, Jannis; Drechsler, Rolf: Effects of Cell Shapes on the Routability of Digital Microfluidic Biochips. In: Design, Automation and Test in Europe. DATE, S. 1627–1630, 2017.

[SHC06] Su, Fei; Hwang, William; Chakrabarty, Krishnendu: Droplet routing in the synthesis of digital microfluidic biochips. In: Design, Automation and Test in Europe. Jgg. 1. IEEE, S. 1–6, 2006.

[Wi15] Wille, Robert; Keszocze, Oliver; Boehnisch, Tobias; Kroker, Alexander; Drechsler, Rolf: Scalable One-Pass Synthesis for Digital Microfluidic Biochips. Design & Test, 32(6):41–50, 12 2015.

[XC06] Xu, Tao; Chakrabarty, Krishnendu: Droplet-Trace-Based Array Partitioning and a Pin Assignment Algorithm for the Automated Design of Digital Microfluidic Biochips. In: International Conference on Hardware/Software Codesign and System Synthesis. ACM, S. 112–117, 2006.

**Oliver Keszöcze** studierte Technomathematik (Diplom) und Informatik (B.Sc.) an der Universität Bremen. Nach dem erfolgreichen Abschluss der Studiengänge 2011 arbeitete er ein Jahr als Softwareentwickler. Seit 2012 ist er als wissenschaftlicher Mitarbeiter an der Universität Bremen in der Arbeitsgruppe Rechnerarchitektur in Forschung und Lehre tätig. Zusätzlich arbeitet er seit 2014 als Researcher beim Deutschen Forschungszentrum für künstliche Intelligenz (DFKI) im Bereich Cyber-Physical Systems.

Das Forschungsinteresse von Herrn Keszöcze ist im Bereich neuartiger Technologien mit Fokus auf dem Entwurf von Biochips angesiedelt. Dies war auch das Thema seiner Promotion, welche er 2017 mit dem Prädikat „summa cum laude“ abschloss. Weitere Forschungsaspekte sind, unter anderem, Entwurf von Quantencomputern, Optical Computing und SAT.

# Varianten der Graph Laplacian mit Anwendungen im Maschinellen Lernen[1]

Sven Kurras[2]

**Abstract:** Graphen lassen sich auf viele Arten als Matrix repräsentieren, zum Beispiel anhand ihrer Adjanzenzmatrix. Die Spektrale Graphentheorie untersucht Graphen anhand der Eigenwerte und Eigenvektoren ihrer Matrizen. Dabei ist vor allem die Graph Laplacian Matrix von Bedeutung, aber es gibt derer viele Varianten. Die hier vorgestellte Dissertation erforscht solche Varianten und beweist neuartige Zusammenhänge zu graphentheoretischen Eigenschaften. Neben der Theorie liegt der Schwerpunkt dabei auf einer umfassenden intuitiven Aufbereitung der mathematischen Ergebnisse, um ein möglichst breites Verständnis zu ermöglichen. Insbesondere werden hierfür vielfältige Anwendungen im Maschinellen Lernen herausgearbeitet. Zudem wird ein neuartiger, auf dem kleinsten Eigenvektor der Signed Laplacian Matrix basierender Correlation-Clustering-Algorithmus entwickelt und in einer umfassenden Praxisanwendung implementiert, welche in Echtzeit unfaires Gruppenverhalten in einem Multiplayer-Online-Spiel detektiert.

## 1 Einführung

Ein Graph $G = (V, E)$ besteht aus einer Menge $V$ beliebiger Objekte ("Knoten"), sowie einer Menge $E$ paarweiser Verbindungen ("Kanten") zwischen Objekten. Solche Graphen finden sich überall. Zum Beispiel repräsentieren die Hyperlinks zwischen Internetseiten einen Graphen, dessen strukturelle Eigenschaften insbesondere für Suchmaschinenbetreiber interessant sind. Weitere Beispiele für Graphen sind soziale Netzwerke, der öffentliche Nahverkehr, Proteine, Finanztransaktionen, Familienstammbäume und der erstaunliche Collatz-Graph. Graphen können für die formale Analyse durch eine reellwertige Matrix repräsentiert werden, zum Beispiel durch die wohlbekannte gewichtete Adjazenzmatrix $W$. Darüber hinaus existieren viele weitere Graphmatrizen, etwa die Diagonalmatrix $D$, welche entlang ihrer Hauptdiagonalen die Grade aller Knoten auflistet, sowie die der Dissertation namensgebende Laplace-Matrix $L := D - W$. Viele strukturelle Eigenschaften von Graphen spiegeln sich auf die ein oder andere Weise in algebraischen Eigenschaften ihrer Graphmatrizen wider. Diese grundlegende Dualität ermöglicht das Studium von Graphen durch Anwendung sämtlicher Resultate der Linearen Algebra auf Graphmatrizen. Beispielsweise ist es leicht einzusehen, dass ein gerichteter Graph genau dann kreisfrei ist, wenn es ein $k$ gibt, für welches $W^k$ die Nullmatrix liefert. Es ist ungleich schwieriger zu sehen, dass für jeden ungewichteten Graphen die Anzahl seiner verschiedenen minimalen Spannbäume gleich dem Produkt aller positiven Eigenwerte von $L$ ist, dividiert durch die Anzahl seiner Knoten. Dieses faszinierende Ergebnis von Kirchhoff aus dem Jahr 1847 begründete die Spektrale Graphentheorie, als das

[1] Englischer Titel der Dissertation: "Variants of the Graph Laplacian with Applications in Machine Learning"

[2] Universität Hamburg, sven.kurras@uni-hamburg.de

Studium von Graphen anhand der Eigenwerte und Eigenvektoren ihrer Graphmatrizen. Dabei ist vor allem die Laplace-Matrix $L$ von Bedeutung, aber es gibt derer viele Varianten, zum Beispiel die normalisierte Laplacian $\mathcal{L} := D^{-1/2}LD^{-1/2}$, die vorzeichenlose Laplacian und die Diplacian. Varianten der Laplacian basieren meist auf einer "syntaktisch kleinen" Änderung von $L$, etwa $D+W$ anstelle von $D-W$. Allerdings ändern solche Modifikationen grundlegend die in den Eigenwerten und Eigenvektoren codierte Information. Somit können sich gänzlich neuartige spektrale Zusammenhänge zu Grapheigenschaften ergeben. Die Dissertation untersucht derartige Varianten der Laplacian. Dabei offenbart sich stets die erstaunliche Komplexität, die eine "einfache" Modifikation auf das Spektrum einer Matrix und auf die Interpretation korrespondierender Grapheigenschaften ausüben kann.

Den drei Hauptteilen der Dissertation ist im Folgenden je ein eigener Abschnitt gewidmet.

## 2 f-adjusted Laplacian

In der Dissertation wird zunächst eine als "**f**-adjusting" bezeichnete Graphmodifikation eines ungerichteten Graphen eingeführt (d.h., eine Abbildung auf der Menge aller symmetrischen Adjazenzmatrizen). Diese verändert sämtliche existierenden Kantengewichte und kann dem Graphen Schleifen hinzufügen oder entfernen. Diese Modifikation wird dann auf zwei Arten interpretiert: Eine *algebraische* Interpretation zeigt, dass **f**-adjusting, durch die Brille der normalisierten Laplacian betrachtet, eine sehr "natürliche" Modifikation eines Graphen ist. Eine *geometrische* Interpretation, im Kontext zufälliger geometrischer Nachbarschaftsgraphen (RGNG), öffnet darüber hinaus die Tür zu vielfältigen Anwendungen des Maschinellen Lernens, von denen abschließend einige skizziert werden.

Für einen beliebigen positiven Vektor $\mathbf{f}$ bezeichne $F := \operatorname{diag}(\mathbf{f})$ die zugehörige Diagonalmatrix. Weiterhin bezeichne $\mathbf{d} := W\mathbf{1}$ den Gradvektor (Zeilensummen) von $W$, somit gilt $D = \operatorname{diag}(\mathbf{d})$. Wir definieren nun drei durch $\mathbf{f}$ parametrisierte Graphmodifikationen:

**f-scaling**: $W \mapsto \tilde{W}_\mathbf{f} := F^{1/2}D^{-1/2} \cdot W \cdot D^{-1/2}F^{1/2}$ multipliziert jeden Eintrag $w_{ij}$ in $W$ mit $\sqrt{f_i f_j}/\sqrt{d_i d_j}$, und repräsentiert den Versuch, im Graphen durch Umgewichtung der Kanten ungefähr den neuen Gradvektor $\mathbf{f}$ zu erhalten.

**f-selflooping**: $W \mapsto W^\circ_\mathbf{f} := W - D + F$ modifiziert die Schleifen aller Knoten derart, dass der resultierende Gradvektor exakt gleich $\mathbf{f}$ ist. Hierdurch können auch negative Schleifengewichte erzeugt werden, jeder Knotengrad bleibt aber positiv (da $f_i > 0$ für alle $i \in V$).

**f-adjusting**: $W \mapsto \overline{W}_\mathbf{f}$ ist definiert über die Hintereinanderausführung von zunächst **f**-scaling, gefolgt von **f**-selflooping.

Auf den ersten Blick erscheint die Definition von **f**-adjusting sehr artifiziell. Allerdings offenbart sich eine erstaunliche Klarheit beim Betrachten der normalisierten Laplacian. In der Dissertation wird gezeigt, dass es eine durch $\mathbf{f}$ determinierte Diagonalmatrix $Z_\mathbf{f}$ gibt, so dass $\mathcal{L}(\overline{W}_\mathbf{f}) = Z_\mathbf{f} - D^{-1/2}WD^{-1/2}$ gilt. Dies offenbart, dass jedes **f**-adjusting als eine Modifikation der originalen normalisierten Laplacian $\mathcal{L}(W) = I - D^{-1/2}WD^{-1/2}$ entlang der Hauptdiagonalen aufgefasst werden kann. Dabei wird lediglich $I$ durch $Z_\mathbf{f}$ ersetzt. Zudem

repräsentiert $Z_{\mathbf{f}} = \operatorname{diag}(\tilde{\mathbf{f}}/\mathbf{f})$ den "relativen Fehler" des direkt nach dem **f**-scaling erhaltenen Gradvektors $\tilde{\mathbf{f}} \neq \mathbf{f}$. Diese Charakterisierung wird in der Dissertation durch das Zeigen der Umkehrrichtung komplettiert: *Jede* Modifikation der Hauptdiagonalen von $\mathcal{L}(W)$, welche wieder eine normalisierte Laplacian liefert, entspricht genau einem **f**-adjusting von $W$. Die zugrundeliegenden Beweise verwenden neben der positiven semi-Definitheit von $\mathcal{L}$ insbesondere Eindeutigkeitsaussagen des Perron-Frobenius-Theorems für irreduzible Matrizen. Dabei werden zudem eine Reihe bekannter Eigenschaften der normalisierten Laplacian auf Graphen mit negativen Schleifengewichten erweitert.

Zufällige geometrische Nachbarschaftsgraphen (RGNGs) sind ein gängiges Modell im unüberwachten und semi-überwachten graphbasierten Lernen. Sie bieten zum Beispiel das mathematische Fundament für bekannte Clustering-Verfahren wie DBSCAN, OPTICS, Mean Shift und Label Propagation. Ihnen liegt folgende Konstruktion zugrunde: Eine Stichprobe der Größe $n$ sei unabhängig und identisch gemäß einer Wahrscheinlichkeitsdichte $p$ auf $\Omega \subset \mathbb{R}^d$ verteilt. Benachbarte Samplepunkte werden nun anhand einer geeigneten Definition von "Nachbarschaft" mit Kanten verbunden, beispielsweise wenn sie im Abstand höchstens $r$ zueinander liegen. Für die formale Analyse sind all diese Parameter zugänglich. In Praxisanwendungen ist hingegen oftmals nur die Adjanzenzmatrix $W$ dieses Graphen gegeben, nicht aber die zugrundeliegende Samplingdichte $p$, ihr Support $\Omega$, die Dimension $d$ oder die Koordinaten der Samplepunkte. Trotzdem spiegeln sich allein in $W$, sowie in daraus ableitbaren Varianten der Laplacian, interessante Eigenschaften der zugrundeliegenden Dichte wieder. Insbesondere verhält sich der Gradvektor **d** in bestimmter Weise proportional zur Dichte $p$. Genauer: der Grad eines Knotens ist bis auf einen globalen Normalisierungsfaktor für $n \to \infty$ ein konsistenter Dichteschätzer der Dichte $p$. Darüber hinaus zeigt [MvLH13], dass sich für $S \subseteq V$ auch Graph-Volumina $\operatorname{vol}_G(S) = \sum_{i \in S} d_i$ und Graph-Schnitte $\operatorname{cut}_G(S) = \sum_{i \in S, j \notin S} w_{ij}$ eines solchen Graphen in einen Zusammenhang mit korrespondierenden Bereichsintegralen der zugrundeliegenden Dichte bringen lassen. Anders als die Knotengrade korrespondieren $\operatorname{vol}_G$ und $\operatorname{cut}_G$ erstaunlicherweise aber nicht zu ihren kontinuierlichen Pendants mit Bezug auf $p$, sondern mit Bezug auf die quadrierte Dichte $p^2$. Diese Anomalie beeinflusst die Interpretation sämtlicher graphbasierter Lernverfahren, welche Volumina und Schnitte des Graphen verwenden, insbesondere Spektrales Clustering.

In der Dissertation wird nun der Frage nachgegangen, wie sich **f**-adjusting eines RGNG mit Blick auf die zugrundeliegende Wahrscheinlichkeitsdichte $p$ interpretieren lässt. Die Kerneinsicht ist, dass in der Anomalie einer der beiden Faktoren $p$ von der unveränderbaren Positionsverteilung der Samplepunkte herrührt, der andere Faktor $p$ jedoch von den Kantengewichten, welche sich in Praxisanwendungen durch Umgewichtung beeinflussen lassen. Durch eine Separierung dieser zwei Effekte in der originalen Beweisführung von [MvLH13] wird gezeigt, dass sich Volumina und Schnitte des **f**-adjusted RGNG auf die kontinuierlichen Pendants von $p \cdot f$ beziehen. Dabei ist $f$ eine Funktion auf $\Omega$, deren Auswertung an den Samplepunkten genau die Einträge in **f** sind. Somit lassen sich Volumina und Schnitte eines **f**-adjusted RGNG aufgrund $p \cdot f$ als eine Transformation der zugrundeliegenden Samplingdichte auffassen. Da der Gradvektor **d** des originalen Graphen zu $p$ korrespondiert, ergeben sich folgende Anwendungsmöglichkeiten:

Der **d**-adjusted Graph ist der unveränderte Graph selbst. Volumina und Schnitte beziehen sich darin auf $p \cdot p$. Im **1**-adjusted Graph entsprechen Volumina und Schnitte hingegen $p \cdot 1$, also $p$ selbst. Dies bietet erstmals eine Korrektor der Anomalie, und erlaubt zum Beispiel Spektrales Clustering mit Bezug auf die originale Dichte $p$ anstelle von $p^2$. Allgemein bezieht sich der $\mathbf{d}^q$-adjusted Graph auf $p^{1+q}$, erlaubt also das Studium auch nichtganzzahliger Intensivierungen von $p$. Weiterhin bezieht sich der $\mathbf{d}^{-1}$-adjusted Graph auf die Gleichverteilung auf $\Omega$. Dies erlaubt es, im Graph-Clustering die rein geometrische Struktur von $\Omega$ zu erfassen, ohne von der Samplingdichte beeinflusst zu sein. Ein solches "Herausrechnen" einer unerwünschten nicht-uniformen Samplingdichte wird in der Dissertation am Beispiel einer Bildsegmentierung im Falle nicht-uniform gesampelter Pixelpositionen gezeigt. Eine weitere Anwendung ist die Multiskalen-Clusteranalyse, indem die Einträge von **f** für jeden Knoten anhand der gemittelten Knotengrade innerhalb einer größeren Nachbarschaft bestimmt werden. Auch eignet sich **f**, um im semi-überwachten Lernen Informationen über die Distanz zum nächstliegenden gelabelten Knoten einzubringen. Diese knotenlokale Information kann dann per **f**-adjusting mittels modifizierter Volumina und Schnitte für Algorithmen wie Label Propagation sichtbar gemacht werden.

Zusammengefasst bietet **f**-adjusting, respektive die Modifikation der Hauptdiagonalen der normalisierten Laplacian, ein praktisch und intuitiv einsetzbares Tool, um RGNGs in neuer Art und Weise einzusetzen. Die Ergebnisse dieses Abschnitts wurden auf der 31. International Conference on Machine Learning (ICML) veröffentlicht [KvB14].

## 3 Symmetric Iterative Proportional Fitting

Das im vorigen Abschnitt eingeführte **f**-scaling zielte darauf ab, im modifizierten Graphen $\tilde{W}_{\mathbf{f}} := F^{1/2}D^{-1/2} \cdot W \cdot D^{-1/2}F^{1/2}$ den neuen Gradvektor **f** anzunehmen. Dies gelingt jedoch nur näherungsweise, wobei die Residuen stark von strukturellen Eigenschaften des Graphen abhängen. Die in der Dissertation verfolgte Idee ist nun, **f**-scaling immer wieder erneut auf die jeweils zuvor erhaltene Näherung anzuwenden, um eventuell zu einer Limit-Matrix mit Gradvektor **f** zu konvergieren. Experimente zeigen, dass dies tatsächlich oftmals gelingt, aber manchmal auch nicht. Die theoretische Analyse offenbart Parallelen des iterierten **f**-scaling zum sogenannten Iterative Proportional Fitting (IPF). Hierbei ist eine nicht-negative Matrix $W \in \mathbb{R}^{m \times n}_{\geq 0}$ mit beliebigen positiven Zeilen- und Spaltensummen gegeben, sowie zwei Vektoren **r** und **c** neu zu erreichender positiver Zeilen- und Spaltensummen. Gesucht ist eine nicht-negative Matrix $\hat{W}$, welche gleichzeitig die gewünschten Zeilensummen $\mathbf{r} = \hat{W}\mathbf{1}$ und Spaltensummen $\mathbf{c} = \hat{W}^T\mathbf{1}$ annimmt, und dabei die Nulleinträge von $W$ erhält ($w_{ij} = 0 \Rightarrow \hat{w}_{ij} = 0$). Die Schwierigkeit liegt dabei in der Gleichzeitigkeit des Erreichens aller Anforderungen, denn offensichtlich liefert die reine Zeilenskalierung $\mathrm{diag}(\mathbf{r}) \cdot \mathrm{diag}(W\mathbf{1})^{-1} \cdot W$ problemlos die gewünschten Zeilensummen, ebenso die Spaltenskalierung $W \cdot \mathrm{diag}(W^T\mathbf{1}) \cdot \mathrm{diag}(\mathbf{c})$ die gewünschten Spaltensummen, jeweils unter exaktem Erhalt der Nulleinträge von $W$. Die Idee von IPF ist nun, diese Zeilen- und Spaltenskalierung fortwährend *abwechselnd* anzuwenden, um so eventuell zu einer geeigneten Limit-Matrix zu konvergieren. Das heißt, IPF bezeichnet die folgende rekursiv definierte Folge

von Matrizen $(W_0, W_1, W_2, \ldots)$:

$$W_0 := W, \qquad W_{k+1} := \begin{cases} \mathrm{diag}(\mathbf{r}) \cdot \mathrm{diag}(W_k \mathbf{1}) \cdot W_k & , \quad k \text{ gerade} \\ W_k \cdot \mathrm{diag}(W_k^T \mathbf{1}) \cdot \mathrm{diag}(\mathbf{c}) & , \quad k \text{ ungerade} \end{cases}$$

Die Vermutung ist, dass der Grenzwert $W_\infty := \lim_{k\to\infty} W_k$ existiert und $W_\infty$ gleichzeitig die Zeilensummen $\mathbf{r}$ und Spaltensummen $\mathbf{c}$ aufweist. Diese Fragestellung wird in der Literatur bereits seit den 1930er Jahren in vielfältiger Weise untersucht. Dabei zeigt sich, dass insbesondere das Zulassen von Nulleinträgen in $W$ die Komplexität erhöht, da nun die Konvergenz nur noch für bestimmte Wahlen von $\mathbf{r}$ und $\mathbf{c}$ gilt, abhängig von der konkreten Struktur der Nulleinträge in $W$. Insbesondere kann die Limit-Matrix $W_\infty$ auch neue Nulleinträge aufweisen, die in keinem $W_k$ vorliegen.

Bregman [Br67] untersucht dieses Problem allgemeiner anhand iterativer Projektionen auf konvexe Mengen. Sei $\mathcal{R}$ die Menge aller nicht-negativen $m \times n$-Matrizen mit Zeilensummen $\mathbf{r}$ und $\mathcal{C}$ die Menge derjenigen mit Spaltensummen $\mathbf{c}$. Es ist leicht einzusehen, dass $\mathcal{R}$ und $\mathcal{C}$ konvexe Mengen sind, und dass $\mathcal{R} \cap \mathcal{C}$ nicht-leer ist. Darüber hinaus kann gezeigt werden, dass die oben eingeführte Zeilenskalierung die sogenannte RE-Projektion von $W_k$ auf $\mathcal{R}$ ist. RE steht dabei für Relative Entropie (auch als Kullback-Leibler-Divergenz bekannt) und ist ein in der Informationstheorie gängiges Distanzmaß, sowie ein Spezialfall der allgemeinen Familie von Bregman-Divergenzen. Die RE-Projektion von $W_k$ auf $\mathcal{R}$ liefert also dasjenige Element aus $\mathcal{R}$, welches die Relative Entropie zu $W_k$ minimiert. In seiner Arbeit zeigt Bregman, dass immer wenn der Schnitt mehrerer konvexer Mengen nicht-leer ist, die zyklisch iterierte Bregman-Projektion (auf jede Menge einzeln nacheinander ausgeführt) letztlich zu einem Element im Schnitt aller Mengen konvergiert. Für diesen Grenzwert folgt jedoch *nicht*, dass er optimal ist, d.h., der Grenzwert entspricht im Allgemeinen *nicht* der direkten Projektion des Anfangspunktes auf den Schnitt.

Da IPF also die iterierte RE-Projektion auf $\mathcal{R}$ und $\mathcal{C}$ ist, folgt aus obigen Ergebnissen die Konvergenz von IPF zu einer Matrix der gewünschten Zeilen-/Spaltensummen und Nulleinträge, wann immer auch nur irgendeine solche Matrix existiert. Ergänzend dazu zeigt [Cs75] mit maßtheoretischen Argumenten, dass diese Lösungsmatrix (unabhängig davon, ob sie per IPF gefunden werden kann) tatsächlich die RE-Projektion von $W$ auf $\mathcal{R} \cap \mathcal{C}$ sein muss. Diese beiden separaten Ergebnisse zusammengenommen zeigen also die Konvergenz und RE-Optimalität von IPF. Ein Schritt hin zu einem einheitlichen Beweis für Konvergenz und RE-Optimalität erfolgt in [BL00]. Dort wird die Idee des Dykstra-Algorithmus[2] aufgegriffen und die iterierte Projektion so abgeändert, dass jedes Urbild vor seiner Projektion zunächst um einen "Reflexionsterm" verschoben wird. Der Grenzwert der so modifizierten Folge ist dann im Falle eines nicht-leeren Schnitts nicht nur irgendeine Lösung, sondern genau die Projektion des Ausgangspunkts. Weiterhin wird gezeigt, dass die Reflexionsterme auch weggelassen werden können, ohne Konvergenz oder Optimalität zu beeinflussen, falls alle Mengen affin sind. Allerdings sind weder $\mathcal{R}$ noch $\mathcal{C}$ affin. In der Dissertation wird nun bewiesen, dass die Reflexionsterme sogar dann ohne Auswirkungen weggelassen werden können, wenn lediglich *lokale* Affinität gilt. Tatsäch-

[2] der Dykstra-Algorithmus von Richard L. Dykstra ist nicht zu verwechseln mit dem deutlich bekannteren Dijkstra-Algorithmus von Edsger W. Dijkstra

lich ist bei IPF jedes $W_k$ lokal affin, was man daran sehen kann, dass jedes $W_k$ exakt dieselben Nulleinträge wie $W$ hat. Dies liefert den fehlenden Baustein, um IPF vollständig per geometrischer Intuition zu erfassen. Ein wesentlicher Beitrag der Dissertation zu diesem Themengebiet ist die folgende vereinheitlichte geometrische Intuition für Konvergenz *und* RE-Optimalität von IPF:

*IPF ist der Dykstra-Algorithmus mit RE-Projektionen unter Auslassung seiner Reflexionsterme, was aufgrund lokaler Affinität nichts an seinem Konvergenzverhalten und der RE-Optimalität des Grenzwertes ändert.*

Darüber hinaus zeigt die Dissertation für symmetrisches $W$ und $\mathbf{r} = \mathbf{c} = \mathbf{f}$ die Konvergenz und RE-Optimalität des iterierten **f**-scalings unter exakt denselben Bedingungen wie IPF. Daher wird iteriertes **f**-scaling als symmetrisches IPF bezeichnet (SIPF). Geometrisch kann SIPF so verstanden werden, dass $W_k$ in jedem Schritt *simultan* auf $\mathcal{R}$ und $\mathcal{C}$ projiziert wird, um dann beide Projektionen per geometrischem Mittel in $W_{k+1}$ zusammenzuführen. Der Vorteil von SIPF gegenüber IPF ist, dass die Iteration stets entlang symmetrischer Matrizen erfolgt, sich also jederzeit als ungerichteter Graph auffassen lässt. In der Dissertation werden zudem Konvergenzkriterien für SIPF herausgearbeitet, die sich aus $W$ und **f** anhand von Grapheigenschaften festmachen lassen. Dazu wird der Begriff der "strikten/nicht-striken schwachen **f**-Expansion" eingeführt, welche eine Gewichtszunahme beim Erweitern jeder Teilmenge auf ihre Nachbarschaft fordert. Anhand dessen werden nun drei Fälle für den Grenzwert der SIPF-Sequenz charakterisiert:

**faktorisiert lösbar**: $W$ ist ein strikter schwacher **f**-Expander $\iff$ $W_\infty$ liefert den Gradvektor **f** auf derselben Kantenmenge wie $W$ $\iff$ es gilt $W_\infty = T \cdot W \cdot T$ für eine positive Diagonalmatrix $T$

**nur nicht-faktorisiert lösbar**: $W$ ist ein nicht-strikter schwacher **f**-Expander $\iff$ $W_\infty$ liefert den Gradvektor **f** auf einer echten Teilmenge der Kanten von $W$, während dies auf der Menge aller Kanten von $W$ unmöglich ist $\iff$ es gilt $W_\infty = \lim_{k\to\infty} T_k \cdot W \cdot T_k$ für geeignete Diagonalmatrizen $T_k$, aber $W_\infty \neq T \cdot W \cdot T$ für jede Diagonalmatrix $T$.

**unlösbar**: $G$ erfüllt nicht die schwache **f**-Expansion $\iff$ für keine Teilmenge der Kanten von $W$ können Kantengewichte gefunden werden, die den Gradvektor **f** liefern.

Aus den Resultaten dieses Abschnitts folgt insbesondere die grundsätzliche Einsicht, dass die normalisierte Adjazenzmatrix $D^{-1/2}WD^{-1/2}$ dem ersten Schritt der SIPF-Sequenz für $\mathbf{f} = \mathbf{1}$ entspricht und damit der Approximation einer Matrix mit Zeilen- und Spaltensummen gleich $\mathbf{1}$, genauer, derjenigen doppelt-stochastischen Matrix welche $W$ bezüglich Relativer Entropie am nächsten liegt. Somit kann überall wo die normalisierte Adjazenzmatrix zum Einsatz kommt, auch $W_\infty$ eine interessante Alternative sein, insbesondere wenn $W_\infty = T \cdot W \cdot T$ faktorisiert. Dieser Ansatz wird in ähnlicher Weise in [Im12] verfolgt, um Verzerrungen in der Messung der Interaktionshäufigkeiten zwischen Abschnitten eines Genoms zu eliminieren. Die Ergebnisse dieses Abschnitts wurden auf der 18. International Conference on Artificial Intelligence and Statistics (AISTATS) veröffentlicht [Ku15].

## 4 Spektrales Correlation-Clustering

Ähnlichkeitsgraphen nutzen positive Kantengewichte, um die Ähnlichkeit zwischen zwei Knoten zu quantifizieren. Der minimale Schnitt MinCut$(S,\overline{S})$ ist in solchen Graphen effizient berechenbar. Da er jedoch stark dazu neigt vereinzelte Knoten abzutrennen, ist er für Clustering-Anwendungen ungeeignet. Interessanter sind daher minimale *balancierte* Schnitte, welche durch einen zusätzlichen Straffaktor zu starke Ungleichgewichte zwischen $S$ und seinem Komplement $\overline{S}$ vermeiden. Durch die Balancierung werden die Optimierungsprobleme jedoch NP-hart. Der Erfolg von Spektralem Clustering rührt daher, dass sich eine fraktionale Relaxierung dieser Optimierungsprobleme anhand der Berechnung bestimmer Eigenvektoren, z.B. denen der Laplacian $L$ für Ratio-Cut, effizient lösen lässt. Der Schlüssel hierfür ist die Rayleigh-Quotient-Charakterisierung (RQC). Diese stellt die Eigenvektoren (EV) einer symmetrischen Matrix als Minimierer eines Optimierungsproblems dar. Für $L$ erhält man so den Zusammenhang $EV(L) \Leftrightarrow \mathrm{RQC}(L) \overset{relax}{\rightsquigarrow} \mathrm{MinRatioCut}$.

In Anwendungen mit zusätzlicher Unähnlichkeitsinformation stellt sich die Frage, wie man diese sinnvoll in Clustering-Algorithmen berücksichtigen kann. Zwar lassen sich Unähnlichkeiten als sehr kleine positive Ähnlichkeitswerte ($w_{ij} \approx 0$) darstellen, aber die meisten Algorithmen ignorieren sie dann einfach, gleich so als wenn keine Kante ($w_{ij} = 0$) vorläge. Auf diese Weise eingebracht erzwingen starke Unähnlichkeiten also kein anderes Ergebnis. Dieses Problem lässt sich für Spektrales Clustering eines zusammenhängenden Graphen direkt an RQC($L$) ablesen. Der zweitkleinste Eigenvektor von $L$ minimiert darin $\sum_{ij} w_{ij}(f_i - f_j)^2$ über alle Vektoren $\mathbf{f} \neq \mathbf{0}$, unter der Nebenbedingung $\sum_i f_i = 0$ (ohne die Nebenbedingung erhielte man sofort die Lösung $f_k = 1$ für alle $k$, also $\mathbf{f} = \mathbf{1}$, was genau dem "trivialen" kleinsten Eigenvektor von $L$ entspricht). Aufgrund der Nebenbedingung erhält man positive und negative Einträge in $\mathbf{f}$, aus deren Vorzeichen sich dann der approximierte MinRatioCut$(S,\overline{S})$ ablesen lässt. Die Summendarstellung zeigt, dass je größer die Ähnlichkeit $w_{ij} \gg 0$ ist, sich die Einträge $f_i$ und $f_j$ im Eigenvektor stärker gleichen werden, um den Summanden insgesamt klein zu halten. Werte $w_{ij} \approx 0$ führen jedoch nicht verstärkt dazu, dass die Einträge $f_i$ und $f_j$ in Bereiche verschiedener Vorzeichen auseinandergeschoben werden. Stattdessen werden sie ebenso wie $w_{ij} = 0$ einfach als Freiheiten verstanden, um die eigentliche Ähnlichkeitsoptimierung durchzuführen.

Die Dissertation untersucht nun, wie sich Korrelationsmaße in spektralen Methoden berücksichtigen lassen. Eine positive Korrelation $x \gg 0$ repräsentiert eine starke klare Ähnlichkeit, $x \ll 0$ eine starke klare Unähnlichkeit und $x \approx 0$ leichte Tendenzen bzw. unklare oder nicht vorhandene Information. Correlation-Clustering wurde in der Literatur bereits eingehend untersucht. Dabei zeigt sich ein interessanter Unterschied zum Ähnlichkeits-Clustering: In einem korrelationsgewichteten Graphen ist der MinCut nicht länger uninteressant, sondern in höchstem Maße der geeignete Kandidat für die Clustersuche. Der MinCut wird weiterhin nur wenige positiv gewichtete Kanten schneiden, aber er schneidet nun vehement entlang negativ gewichteter Kanten, solange dies das Gesamtschnittgewicht senkt. Insbesondere ist der MinCut nun im Allgemeinen negativ. Das in der Literatur verbreitete CC-Funktional beschränkt sich dabei nicht auf einen einfachen Schnitt, sondern entspricht der Minimierung des $k$-MinCut für unbestimmtes $k$, wobei $k$ explizit auch 1 sein darf (wenn alle Knoten untereinander eher positiv korreliert sind, dann ist ganz $V$

tatsächlich der sinnvollste Correlation-Cluster). Diese erhöhte Ausdruckskraft des minimalen Schnitts reellwertiger Graphen geht allerdings damit einher, dass dieser nicht mehr effizient zu berechnen ist, sondern NP-hart (selbst für fest gewähltes $k > 1$).

Es existieren zahlreiche nicht-spektrale Heuristiken zur Optimierung des CC-Funktionals. Zudem gibt es vereinzelte spektrale Ansätze, welche das Konzept balancierter Schnitte auf Korrelationsgewichte übertragen. An dieser Stelle verfolgt die Dissertation einen gänzlich neuen Ansatz: Die nicht-spektralen Ansätze zeigen, dass im Correlation-Clustering keine zusätzliche Balancierung benötigt wird. Die Clustergrenzen sollen sich ausschließlich anhand der Korrelationsgewichte orientieren, also am $k$-MinCut. In der Dissertation wird erstmals eine unbalancierte spektrale Clustering-Methode entwickelt. Hierfür wird die reelle Adjazenzmatrix $W$ zerlegt in $W = A - R$, wobei $A$ die positiven Einträge von $W$ und $R$ die Absolutbeträge der negativen Einträge enthält. Dann wird das OptMinCut-Problem definiert als $k$-MinCut für $k \in \{1,2\}$ und gezeigt, dass sich dieses darstellen lässt als die Minimierung von $\sum_{ij} a_{ij}(f_i - f_j)^2 + \sum_{ij} r_{ij}(f_i + f_j)^2$ über alle $\mathbf{f}$ mit $\|\mathbf{f}\| = \sqrt{n}$. In dieser Summe erzwingen Ähnlichkeiten (große $a_{ij}$) wieder ein Zusammenrücken der Einträge $f_i$ und $f_j$, während Unähnlichkeiten (große $r_{ij}$) nun ein Auseinanderschieben der Einträge hin zu $f_i \approx -f_j$ forcieren. Aus den Vorzeichen in $\mathbf{f}$ erhält man schließlich entweder einen Schnitt oder keinen Schnitt. Letzterer Fall tritt genau dann ein, wenn kein Schnitt mit negativem Gewicht existiert. In der Dissertation wird nun eine Matrix konstruiert, für welche diese Summe exakt ihre RQC ist. Dies gelingt für den kleinsten Eigenvektor der Signed Laplacian $\overline{L} := \overline{D} - W$, wobei die Diagonaleinträge in $\overline{D}$ die Absolutgrade $\overline{d}_i = \sum_j |w_{ij}|$ sind. Insbesondere erhält man somit $L$ für jede nicht-negative Matrix $W$ als Spezialfall von $\overline{L}$. Vor diesem Hintergrund ist es interessant, dass gerade der kleinste Eigenvektor von $\overline{L}$ für das Correlation-Clustering relevant ist, während er im Änlichkeits-Clustering lediglich der triviale $\mathbf{1}$-Vektor ist. Dies ist jetzt dadurch verständlich, dass der kleinste Eigenvektor von $\overline{L}$ immer das OptMinCut-Problem zu lösen versucht. Da im Ähnlichkeits-Clustering anhand $L = \overline{L}$ kein negativer Schnitt existiert, wird OptMinCut dort stets korrekt über den 1-MinCut mit Gewicht 0 gelöst. Seine eigentliche Stärke kann OptMinCut erst im Zusammenspiel mit negativen Kantengewichten entfalten.

Die Dissertation zeigt also, dass $EV(\overline{L}) \Leftrightarrow \mathrm{RQC}(\overline{L}) \overset{relax}{\leftrightsquigarrow} \mathrm{OptMinCut}$ gilt. Ein beachtenswertes Merkmal ist zudem, dass diese spektrale Relaxierung von OptMinCut geometrisch sehr leicht zu verstehen ist: Das NP-harte Problem minimiert die Summe über alle $\mathbf{f} \in \{-1,1\}^n$, also über alle $2^n$ Eckpunkte des $n$-dimensionalen Hypercubes. Die Relaxierung erweitert die Lösungsmenge auf alle $\|\mathbf{f}\| = \sqrt{n}$, also auf die gesamte durch die Ecken des Hypercubes verlaufende Einheitssphäre. Der Fehler durch die abschließende Vorzeichenrundung entspricht geometrisch der Auswahl derjenigen Ecke, welche der fraktionalen Lösung auf der Einheitssphäre am nächsten liegt. Die Dissertation liefert somit die neue Einsicht, dass der kleinste Eigenvektor von $\overline{L}$ als Minimierer einer ganz "naheliegenden" spektralen Relaxierung von OptMinCut angesehen werden kann.

Für die Erweiterung auf $k > 2$ schlägt die Dissertation den folgenden im Clustering üblichen rekursiven Ansatz vor: Im Falle eines Schnitts werden beide Teilgraphen separat rekursiv weiter zerlegt. Dieser Algorithmus wird im Folgenden als SCC-Algorithmus bezeichnet. Eine Besonderheit ist sein automatisches Pruning: Die Rekursion stoppt eigen-

ständig, sobald ein Teilgraph keinen negativen Schnitt mehr beinhaltet. Experimente zeigen, dass der SCC-Algorithmus hierüber tatsächlich in der Lage ist, selbstständig die korrekten Cluster unbekannter Anzahl zu finden. Zudem zeigt ein umfassender quantitativer und qualitativer Vergleich zu nicht-spektralen und spektralen Methoden, dass er mit der aktuell besten Heuristik in puncto Clusterqualität und Rechenaufwand gleichauf ist, und dazu besondere Stärken im Falle dünnbesetzter und verrauschter Graphen zeigt.

Die theoretische Entwicklung des SCC-Algorithmus erfolgte im Kontext einer umfassenden Praxisanwendung, die hier abschließend skizziert werden soll.

Die Firma Sandbox Interactive (Berlin) entwickelt das Fantasy-MMORPG Albion Online, welches mittlerweile im Produktiveinsatz ist. Während der Alpha-Phase sollte im Rahmen eines Proof-of-Concept untersucht werden, ob sich mittels Echtzeit-Clustering-Verfahren ein bestimmtes unerwünschtes Verhalten einiger Spielergruppen automatisch detektieren lässt, so dass es über spielinterne Handicaps geahndet werden kann. Bei diesem Verhalten handelt es sich um trollartiges Zerging und Griefing. Hierbei verbündet sich spontan eine größere Gruppe Teilnehmer, meist Gelegenheitsspieler, um Einzelpersonen oder Kleingruppen umzubringen und bewusst zu verärgern. Grundsätzlich wird zwar eine möglichst große Handlungsfreiheit angestrebt, allerdings kann dies auch zu einem permanenten Vergraulen der zahlenden Stammspieler führen. Im Rahmen der Dissertation wurde hierfür eigenständig eine umfassende Server-Software entwickelt, welche sich in die bestehende Gameserver-Architektur integrieren lässt. Als Eingabe erhält sie sekündlich von allen Gameservern Informationen über alle Interaktionen aller Spieler. Intern pflegt sie damit einen "Weltgraphen", der die positiven Interaktionen (Heilzauber, lange friedliche Nähe, ...) und negativen Interaktionen (Schadenspunkte, Störzauber, ...) speichert und erst im Laufe der Zeit langsam vergisst. Über eine geeignete Abbildung aller Interaktionen auf eine gemeinsame Skala lässt sich der Graph dabei als Korrelationsgraph auffassen. Intern wird dieser Weltgraph nun in Echtzeit, unter Ausnutzung aller verfügbaren Rechenkerne, fortlaufend mittels dem SCC-Algorithmus geclustert. Dies liefert für jeden aktuell in der Spielwelt stattfindenden Kampf ein Clustering der Teilnehmer in ein, zwei oder mehr Teams. Um dabei insbesondere auf "Cheater" zu reagieren, die ihre echte Teamzugehörigkeit über gelegentliches "friendly fire" oder "enemy healing" zu verstecken versuchen, wurde zudem im SCC-Algorithmus die Pruning-Strategie so modifiziert, dass sie bereits vor Erreichen des Schnittgewichts 0 die weitere rekursive Zerlegung abbricht. Dieser Pruning-Parameter lässt sich intuitiv als Toleranz-Schwellwert für die Spieler-individuelle Zufriedenheit mit dem aktuellen Team-Clustering interpretieren, analog zu Utility-Funktionen in der Spieltheorie. Zur Evaluation wurden durch 30 Alpha-Tester diverse extra konzipierte Kampfszenarien nachgestellt und aufgezeichnet, inklusive der korrekten Teamzugehörigkeit als Ground Truth. Diese Szenarien eigneten sich hervorragend zur Entwicklung, insbesondere durch ein Rekombinierungsmodul, welches diese Szenarien duplizieren, randomisieren und zusammenführen kann, um neue Szenarien aus tausenden Spielern zu generieren. Für die qualitative Analyse wurde zudem ein grafisches Frontend entwickelt, welches interaktive Einsichten in die simulierten und echten Spielsituationen erlaubt, inklusive einer Live-Aufzeichnung, um sich zeitliche Entwicklungen im Clustering und in den Statistiken genauer anzuschauen. Die Simulationen und ein Praxistest vor Ort schließen den Proof-of-Concept ab und zeigen, dass eine derartige Lösung praxistauglich ist.

Die Ergebnisse dieses Abschnitts erfolgten zum Ende der Dissertation und wurden bislang lediglich im Rahmen der Dissertation veröffentlicht. Eine Aufbereitung zur Einreichung bei einer Konferenz wäre aber definitiv zu begrüßen, da die Ergebnisse einen signifikanten Beitrag zum Correlation-Clustering leisten.

## Literaturverzeichnis

[BL00] Bauschke, H. H.; Lewis, A. S.: Dykstra's algorithm with Bregman projections: a convergence proof. Optimization, 48:409–427, 2000.

[Br67] Bregman, L. M.: The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming. USSR Computational Mathematics and Mathematical Physics, 7(3):200–217, 1967.

[Cs75] Csiszar, I.: I-Divergence Geometry of Probability Distributions and Minimization Problems. Annals of Probability, 3(1):146–158, 1975.

[Im12] Imakaev, M.; Fudenberg, G.; McCord, R. P.; Naumova, N.; Goloborodko, A.; Lajoie, B. R.; Dekker, J.; Mirny, L. A.: Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods, 9(10):999–1003, 2012.

[Ku15] Kurras, S.: Symmetric Iterative Proportional Fitting. In: International Conference on Artificial Intelligence and Statistics (AISTATS). S. 526–534, 2015.

[KvB14] Kurras, S.; von Luxburg, U.; Blanchard, G.: The f-Adjusted Graph Laplacian: a Diagonal Modification with a Geometric Interpretation. In: International Conference on Machine Learning (ICML). S. 1530–1538, 2014.

[MvLH13] Maier, M.; von Luxburg, U.; Hein, M.: How the result of graph clustering methods depends on the construction of the graph. ESAIM: Probability and Statistics, 17:370–418, 2013.

**Sven Kurras** wurde am 7. Mai 1978 in Berlin geboren. Er arbeitete bereits einige Jahre als freiberuflicher Softwareentwickler, als er 2003 parallel dazu sein Informatik-Studium an der Universität Paderborn aufnahm. Dort forschte er zudem als wissenschaftliche Hilfskraft am Fraunhofer ENAS/ASE und am PC$^2$ Paderborn Center for Parallel Computing. Er schloss das Master-Studium 2012 mit Auszeichnung ab. Anschließend promovierte er in der Arbeitsgruppe von Ulrike von Luxburg zur Theorie des Maschinellen Lernens an der Universität Hamburg, gefördert durch die DFG-Forschergruppe "Structural Inference in Statistics: Adaptation and Efficiency". Der Abschluss der Promotion erfolgte im März 2017. Ein Jahr zuvor trat er bereits eine Vollzeit-Anstellung als Senior Data Scientist bei der Risk.Ident GmbH in Hamburg an, wo er bis heute graphbasierte Lernverfahren zur Betrugserkennung in den Bereichen E-Commerce und Financial Services entwickelt. Die Nähe zur Forschung bleibt dabei über die regelmäßige Teilnahme an Konferenzen, Seminaren und Workshops, sowie über eigene Gastvorträge erhalten.

Steffen Hölldobler et. al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2018 141

# Verbesserte Algorithmen und Bedingte Untere Schranken für Probleme in Formaler Verifikation und Reaktiver Synthese[1]

Veronika Loitzenbauer[2]

**Abstract:** Die formale Verifikation ist ein Ansatz um Fehler in Computerprogrammen und anderen Systemen automatisiert und systematisch aufzuspüren oder die Erfüllung von gewünschten Eigenschaften zu garantieren. Reaktive Synthese bezeichnet die Erzeugung von Systemen basierend auf einer formalen Spezifikation, so dass sich diese wie gewünscht nach auen hin verhalten. In der Dissertation betrachten wir algorithmische Probleme in formaler Verifikation und reaktiver Synthese aus theoretischer Perspektive. Wir zeigen einerseits Algorithmen mit verbesserten asymptotischen Laufzeitschranken und anderseits bedingte untere Schranken, die zeigen, dass weitere Verbessungen der Laufzeitschranken zu Durchbrüchen im Algorithmendesign für wohlbekannte Probleme führen würden. Wir verbinden damit formale Methoden für Systeme mit Algorithmentheorie und führen neue Techniken und Forschungsrichtungen für Polynomialzeitprobleme in diesem Gebiet ein.

## 1 Einführung

Fehler im Software und Hardware Design können unsere Sicherheit gefährden und hohe Kosten verursachen wenn sie zum Beispiel bei der Steuerung eines Flugzeuges oder im Design eines Prozessors auftreten. Beweisbar fehlerfreie Systeme wären daher sehr wünschenswert. Wir betrachten hier abstrakte Modelle von Systemen. In der Praxis kann so ein System vieles sein, von einer einfachen Verkehrsampel bis zu parallelen Computerprogrammen, Kommunikationsprozessen oder elektronischen Schaltungen. Die Korrektheit von Systemen zu beweisen ist im Allgemeinen unmöglich, jedoch wurden in den letzten Jahrzehnten viele nützliche formale Methoden sowie Tools für die praktische Anwendung entwickelt, die helfen, die Korrektheit von Systemen zu überprüfen. Die *formale Verifikation* ist eine essentielle Komponente im iterativen Design von Systemen wie Mikroprozessoren, Kommunikationsprotokollen und sicherheitsrelevanten Algorithmen geworden. Die *Modellprüfung* ist ein vollautomatisierter Ansatz in der Verifikation um entweder die Korrektheit eines Systems bezüglich bestimmter Eigenschaften zu beweisen oder ein Gegenbeispiel dafür zu finden. Während Verifikation hauptsächlich für iterative Designprozesse verwendet wird, verlangt das *Syntheseproblem* die automatische Generierung von korrekten Systemen aus einer Spezifikation. Die *reaktive Synthese* ist die Synthese von Systemen, die wiederholt mit ihrer Umgebung interagieren und daher *reaktive Systeme* genannt werden. In dieser Arbeit beschäftigen wir uns mit algorithmischen Fragen im Bereich der Modellprüfung und Synthese.

[1] Englischer Titel der Dissertation: "Improved Algorithms and Conditional Lower Bounds for Problems in Formal Verification and Reactive Synthesis" [Lo16]. Siehe [Lo16] für detailierte Literaturverweise. Die Dissertation basiert auf den Publikationen [Ch14b, CHL17, Ch16b, Ch16a, Ch17].

[2] Institut für formale Modelle und Verifikation, Johannes Kepler Universität Linz, veronika.loitzenbauer@jku.at

**Modelle.** Für Verifikation und Synthese werden mathematische *Modelle* der Systeme sowie ein formale Beschreibung ihrer *Spezifikation*, das heißt ihres gewünschten Verhaltens, benötigt. Ein Modell eines Systems beschreibt typischerweise den Kontrollfluss sowie die Interaktion zwischen verschiedenen Teilen des Systems beziehungsweise zwischen dem System und seiner Umgebung, während Details der Implementierung ignoriert werden.

*Endliche gerichtete Graphen* sind ein Modell für nicht-deterministische Systeme. Die Knoten eines Graphen repräsentieren dabei die Zustände des Systems und die Kanten die bergänge zwischen den Zuständen. Die verschiedenen ausgehenden Kanten eines Knotens repräsentieren verschiedene Verhaltensmöglichkeiten des Systems die, zum Beispiel, durch die Nicht-Determiniertheit in parallelen oder verteilten Ausführungen von Prozessen oder, während dem iterativen Design von Systemen, durch unterschiedliche Designmöglichkeiten entstehen können.

*Markow-Entscheidungsprozesse* (*MEPs*) modellieren Systeme, die sowohl Nicht-Determiniertheit als auch zufälliges Verhalten zeigen. MEPs sind Graphen in denen eine Teilmenge der Knoten eine Wahrscheinlichkeitsverteilung über ihre ausgehenden Kanten besitzt. So eine Wahrscheinlichkeitsverteilung kann z.B. Randomisierung in verteilten Systemen, randomisierte Kommunikationsprotokolle, verlustbehaftete Kommunikationskanäle oder experimentelle Daten über die Umgebung des Systems modellieren.

In *Spielgraphen* sind die Knoten zwischen *zwei Spielerinnen* aufgeteilt. Typischerweise repräsentiert eine Spielerin das System und die anderen die Umgebung oder, wie in der Verifikation von Verzweigunszeit Eigenschaften von reaktiven Systemen, eine Spielerin modelliert die existenziellen Quantoren und eine die universellen Quantoren. Zwei-Spieler Spiele auf Graphen sind für viele Probleme in der Verifikation und Synthese wichtig, sowie der Synthese von reaktiven Systemen und der Verifikation von offenen Systemen.

**Spezifikationen.** Der *Automaten-basierte* Ansatz für Modelprüfung und Synthese ist eine anerkannte Methode um das erwünschte Verhalten von Systemen mit Hilfe von $\omega$-*regulären Zielvorgaben* formal zu beschreiben ($\omega$-reguläre Sprachen erweitern reguläre Sprachen auf unendliche Wörter) und kann die meisten häufig verwendeten Eigenschaften in formaler Verifikation und reaktiver Synthese ausdrücken. Jede Zielvorgabe hat eine komplementäre oder *duale* Zielvorgabe. Wenn eine Zielvorgabe das gewünschte Verhalten beschreibt, so beschreibt ihre duale Zielvorgabe das ungewünschte Verhalten. In Spielgraphen haben die beiden Spielerinnen zueinander komplementäre Zielvorgaben.

Die grundlegendsten Eigenschaften eines sicherheitskritischen Systems werden durch *Sicherheitszielvorgaben* beschrieben. Für Sicherheitszielvorgaben wollen wir verifizieren, dass eine gegebene Menge von schlechten Systemzuständen vermieden werden kann. Ein Beispiel für so einen schlechten Systemzustand in einem parallelen System wäre wenn zwei Prozesse gleichzeitig in einem kritischen (d.h. sequentiell zu absolvierenden) Abschnitt des Systems wären. Die duale Zielvorgabe zu einer Sicherheitszielvorgabe ist eine *Erreichbarkeitszielvorgage*, die eine Menge von guten Systemzuständen angibt, die letztendlich erreicht werden soll.

Die Frage ob jede Anfrage eines Prozesses einen kritischen Abschnitt zu durchlaufen schließlich erlaubt wird ist ein Beispiel für eine weitere Art von Eigenschaften, die mit *Büchi*-Zielvorgaben beschrieben wird. Dabei muss eine Menge von guten Zuständen *unendlich oft* erreicht werden. Die dazu duale Zielvorgabe ist die *co-Büchi*-Zielvorgabe, bei der eine Menge von schlechten Zuständen nur endlich oft erreicht werden darf.

*Streett*-Zielvorgaben und ihre dualen *Rabin*-Zielvorgaben können alle $\omega$-regulären Sprachen ausdrücken. Streett-Zielvorgaben entsprechen direkt *starken Fairnessbedingungen*. Ein Scheduler, zum Beispiel, ist *stark fair* wenn jedes Event, das unendlich oft aufgerufen wird, auch unendlich oft eingeplant wird. Eine Streett-Zielvorgabe wird durch $k$ Paare $(L_i, U_i)_{1 \leq i \leq k}$ von Knotenmengen definiert. Zur Erfüllung der Streett-Bedingung muss für alle $k$ Paare ein unendlicher Pfad, der die Menge $L_i$ unendlich oft kreuzt auch die Menge $U_i$ unendlich oft kreuzen. Büchi- und co-Büchi- Zielvorgaben sind wichtige Spezialfälle von Streett- und Rabin Zielvorgaben mit $k = 1$.

*Paritäts*-Zielvorgaben verallgemeineren Büchi-Zielvorgaben und sind ebenso Spezialfälle von sowohl Streett- und Rabin-Zielvorgaben. Umgekehrt können auch Streett- und Rabin-Zielvorgaben in Paritäts-Zielvorgaben umgewandelt werden, jedoch wächst dabei die Modellgröße exponentiell. Die duale Zielvorgabe zu einer Paritäts-Zielvorgabe ist wieder eine Paritäts-Zielvorgabe. Paritäts-Zielvorgaben sind besonders wichtig da sie äquivalent zum modalen $\mu$-Calculus sind, einer der wichtigsten Logiken für die Modellprüfung. Für eine Paritäts-Zielvorgabe werden den Knoten natürliche Zahlen, Prioritäten genannt, zugeordnet. Ein unendlicher Pfad erfüllt eine *gerade* Paritätsbedingung wenn die höchste Priorität, die unendlich oft vorkommt gerade ist, und ansonsten die komplementäre *ungerade* Paritätsbedingung.

Eine andere Verallgemeinerung von Paritäts-Zielvorgaben, die über $\omega$-reguläre Zielvorgaben hinausgeht, sind *Mittelwerts*-Zielvorgaben. Hierfür werden den Kanten des Modells Gewichte zugeordnet und die Zielvorgabe wird bezüglich des Durchschnittsgewichts einer Sequenz von Kanten definiert. Mit solchen *quantitativen Zielvorgaben* kann zum Beispiel der durchschnittlichen Resourcenverbrauch oder die durchschnittliche Verzögerung eines Systems modelliert werden.

**Algorithmische Fragestellungen.** Um über alle möglichen Ausführungen eines Systems argumentieren zu können, betrachten wir unendliche Pfade, die durch das Bewegen eines Markers entlang der Kanten des Modells induziert werden. Die Wahl einer ausgehenden Kante an einem (nicht-zufälligen) Knoten wird als *Strategie* bezeichnet.

In *Graphen* wollen wir für jeden Ausgangsknoten wissen ob es einen unendlichen Pfad gibt, der die Zielvorgabe erfüllt. Die Menge der Ausgangsknoten, für die das der Fall ist, ist die *Gewinnmenge*. Die algorithmische Fragestellung für Graphen ist die Gewinnmenge sowie zugehörige Strategien zu berechnen.

In *MEPs* gibt es zusätzlich Zufallsknoten und an den Zufallsknoten wird die ausgehende Kante anhand der gegebenen Wahrscheinlichkeitsverteilung ausgewählt. In MEPs wollen wir für jeden Ausgangsknoten wissen ob es eine Strategie für die Nicht-Zufallsknoten gibt, so dass die Zielvorgabe mit Wahrscheinlichkeit 1 erfüllt wird. Die Menge der Ausgangs-

knoten, für die das der Fall ist, heißt *quasi-sichere Gewinnmenge*. Die Anforderung die Zielvorgabe mit Wahrscheinlichkeit 1 zu erfüllen wird auch die *qualitative* Analyse von MEPs genannt und ist in der Analyse von randomisierten verteilten Algorithmen üblich.

In *Spielgraphen* werden die unendlichen Pfade durch die Spielzüge der beiden Spielerinnen induziert, wobei jeweils die Eigentümerin eines Knotens entscheidet über welche ausgehende Kante eines Knotens der Marker bewegt wird. Eine *Strategie* einer Spielerin ist eine Funktion, die für jeden Knoten der Spielerin beschreibt wie ein gegebener Pfad fortgesetzt wird. Eine *Gewinnstrategie* stellt sicher dass die Zielvorgabe der Spielerin gegen jede mögliche Strategie der Gegnerin erreicht wird. Die *Gewinnmenge* einer Spielerin ist die Menge aller Ausgangsknoten für die sie eine Gewinnstrategie besitzt. Die Gewinnmengen der zwei Spielerinnen partitionieren den Spielgraphen. Die algorithmische Fragestellung ist es, die Gewinnmengen und die zugehörigen Gewinnstrategien zu bestimmen.

**Symbolische Modellprüfung.** Eine fundamentale Schwierigkeit bei der Modellprüfung ist, dass die Anzahl der Systemzustände, und damit die Anzahl der Knoten des Modells, exponentiell in der Anzahl der Systemvariablen ist. Ein Ansatz um damit umzugehen sind *symbolische* Algorithmen, bei denen die Systemzustände und -übergänge nicht explizit konstruiert werden sondern stattdessen die Vorgänger und Nachfolgerzustände von Mengen von Zuständen bei Bedarf generiert werden. Eine Menge von Zuständen kann hierfür zum Beispiel mit einem binären Entscheidungsdiagramm (BDD) kodiert werden. Für Spielgraphen mit Paritäts-Zielvorgaben (Paritätsspielen) betrachten wir auch symbolische Algorithmen.

**Bedingte untere Schranken.** In dieser Arbeit beschäftigen wir uns hauptsächlich mit Problemen, für die Polynomialzeitalgorithmen bekannt sind, die einzige Ausnahme sind Paritätsspiele. In der Komplexitätstherie gelten Polynomialzeitalgorithmen als effizient. Für sehr große Graphen wie sie in der Modellprüfung auftreten, kann aber zum Beispiel der Unterschied zwischen einem Algorithmus, der in quadratischer Zeit läuft, und einem, der in linearer Zeit läuft, entscheidend für seine praktische Anwendbarkeit sein. Deshalb würden wir gerne auch für Polynomialzeitprobleme Algorithmen mit besseren asymptotischen Laufzeitschranken finden oder zeigen, dass solche Algorithmen nicht existieren können. Unbedingte super-lineare untere Schranken für Polynomialzeitprobleme sind sehr selten. Jedoch wurden in den letzten Jahren *bedingte untere Schranken* für viele verschiedene kombinatorische Probleme gezeigt. Das sind untere Schranken, die auf einer Annahme über die bestmögliche Laufzeitschranke (abgesehen von Termen niedrigerer Ordnung) für ein wohluntersuchtes Problem basieren. Die bedingten unteren Schranken in dieser Arbeit nehmen an dass (A1) es keinen kombinatorischen Algorithmus für die Multiplikation von zwei $n \times n$ booleschen Matrizen mit einer Laufzeit von $O(n^{3-\varepsilon})$ für ein beliebiges $\varepsilon > 0$ gibt oder (A2) es für alle $\varepsilon > 0$ ein $k$ gibt so dass es keinen Algorithmus für das $k$-CNF-SAT Problem gibt der in Zeit $2^{(1-\varepsilon)\cdot n} \cdot \text{poly}(m)$ läuft, wo $n$ die Anzahl der Variablen und $m$ die Anzahl der Klauseln in der $k$-CNF-SAT Formel ist. Kombinatorisch bezeichnet hier die Vermeidung von bestimmten theoretisch schnellen (aber inpraktikablen) Matrixmultiplikationsalgorithmen. Annahme (A2) ist bekannt als Strong Exponential Time Hypothesis (SETH). Diese beiden Annahmen sind bereits für viele bedingte untere Schranken verwendet worden, zum Beispiel im Bereich von dynamischen Graphalgorithmen,

kontext-freier Grammatik und der Verifizierung von Grapheigenschaften erster Ordnung. Bisher ist keine Beziehung zwischen den beiden Annahmen (A1) und (A2) bekannt.

Für einige grundlegende Modellprüfungsprobleme sind die bestbekannten oberen Laufzeitschranken von quadratischer oder kubischer Ordnung und es gibt keine super-linearen unteren Schranken. In dieser Arbeit präsentieren wir mehrere Algorithmen, die verbesserte Laufzeitschranken bieten, und etablieren die ersten (super-linearen) bedingten unteren Schranken für fundamentale Polynomialzeitprobleme in der Modellprüfung und Synthese.

## 2 Forschungsstand

In diesem Abschnitt präsentieren wir einen berblick über die bereits bekannten algorithmischen Ergebnisse für die relevanten Modelle und Zielvorgaben. Wir bezeichnen mit $n$ die Anzahl der Knoten und mit $m$ die Anzahl der Kanten im gegebenen Modell. Für Paritäts-Zielvorgaben bezeichnen wir mit $c$ die Anzahl der den Knoten zugeordneten Prioritäten und für Mittelwerts-Zielvorgaben bezeichnet $W$ das maximale Gewicht einer Kante. Die hier angegebenen Laufzeitschranken berücksichtigen nicht die Abhängigkeit von anderen Eingabeparametern und sind zum Teil zu Gunsten der Lesbarkeit vereinfacht. Der detailierte Forschungsstand findet sich in den entsprechenden Kapiteln der Dissertation [Lo16].

Für *Graphen* kann die Gewinnmenge für Erreichbarkeits-, Sicherheits-, Büchi- und co-Büchi-Zielvorgaben in linearer Zeit ($O(m)$) berechnet werden, für Paritäts-Zielvorgaben in Zeit $O(m \log n)$ [CH14a]. Für Streett-Zielvorgaben gibt es einen $O(m \min(\sqrt{m \log n}, n))$ Zeit Algorithmus [HT96]. Der triviale Algorithmus für Rabin-Zielvorgaben benötigt Zeit $O(mn)$. Für Mittelwerts-Zielvorgaben laufen die besten Algorithmen in Zeit $O(mn)$ [Ka78] und $O(m\sqrt{n}\log(nW))$ [OA92].

Für *MEPs* sind Linearzeitalgorithmen nur für Sicherheits-Zielvorgaben bekannt, für welche das Problem äquivalent zu jenem in Spielgraphen ist. Die quasi-sichere Gewinnmenge für Erreichbarkeits-, Büchi- und co-Büchi-Zielvorgaben kann in Zeit $O(\min(m^{1.5}, n^2))$ berechnet werden [CJH03, CH14a] und mit einem zusätzlichen Faktor von $\log n$ auch für Paritäts-Zielvorgaben [CH14a]. Für Streett- und Rabin-Zielvorgaben folgen Laufzeiten von $O(n \cdot \min(m^{1.5}, n^2))$ aus [CH14a]. Mittelwerts-Zielvorgaben in MEPs können in polynomialer Zeit mit linearer Programmierung und in pseudo-polynomialer Zeit von $O(mnW)$ [FV97] gelöst werden.

In *Spielgraphen* ist die Laufzeit für eine Zielvorgabe und ihre duale Zielvorgabe jeweils die gleiche, da die beiden Zielvorgaben den Zielvorgaben der beiden Spielerinnen entsprechen und die beiden Gewinnmengen die Knoten partitionieren. Für Erreichbarkeits- und Sicherheitszielvorgaben können die Gewinnmengen in linearer Zeit berechnet werden. Für Büchi- und co-Büchi-Zielvorgaben benötigt der bestbekannte Algorithmus Zeit $O(n^2)$ [CH14a]. Für Streett- und Rabin-Zielvorgaben ist das Problem coNP- bzw. NP-vollständig [EJ88]. Für den Spezialfall von 1-Paar Streett und 1-Paar Rabin Zielvorgaben gibt es einen $O(mn)$ Zeit Algorithmus [Ju00]. Paritätsspiele und ihre Verallgemeinerung Mittelwertsspiele sind eine der wenigen "natürlichen" Probleme die in $\mathrm{NP} \cap \mathrm{coNP}$ liegen für die keine Polynomialzeitalgorithmen bekannt sind. Bis vor kurzem waren die

besten bekannten Algorithmen für Paritätsspiele ein $n^{O(\sqrt{n})}$-Zeit [JPZ08] und ein (ca.) $O(m \cdot n^{c/3})$-Zeit [Sc07] Algorithmus, nach Abschluss der Dissertation wurde der erste quasi-polynomial Zeit Algorithmus veröffentlicht [Ca17]. Büchi-Spiele sind Paritätsspiele mit $c = 2$ und Paritätsspiele mit $c = 3$ sind äquivalent zu 1-Paar Streett-Spielen. Die besten Algorithmen für Mittelwertsspiele laufen in pseudo-polynomialer Zeit $O(mnW)$ [Br11] und in randomisierter sub-exponentieller Zeit $O(2^{\sqrt{n \log n}} \log W)$ [BV07].

Paritätsspiele können in $O(n^c)$ vielen *symbolischen Schritten* gelöst werden wenn eine lineare Anzahl von Mengen gespeichert wird [EL86, Zi98] oder mit $O(n^{c/2+1})$ vielen symbolischen Schritten wenn $O(n^{c/2+1})$ viele Mengen verwendet werden [Br97].

## 3 Ergebnisse

Wir fassen nun die Ergebnisse der Dissertation zusammen. Die Laufzeiten hier sind teilweise vereinfacht, die tatsächlichen Verbesserungen hängen noch von weiteren Eingabeparametern ab. Alle Algorithmen für $\omega$-reguläre Zielvorgaben berechnen die (quasi-sicheren) Gewinnmengen und können leicht modifiziert werden so dass auch die zugehörigen Gewinnstrategien innerhalb der gleichen Laufzeitschranken berechnet werden. Die Details befinden sich in den entsprechenden Kapiteln der Dissertation [Lo16].

**Approximationsalgorithmus für das kleineste mittlere Kreisgewicht.** Wir präsentieren den ersten *Approximationsalgorithmus* für *Mittelwerts-Zielvorgaben auf Graphen*, wobei sich die Berechnung der Gewinnmenge auf die Bestimmung des Kreises mit dem kleinstem durchschnittlichem Kantengewicht reduzieren lässt. Dies ist ein grundlegendes graphtheoretisches Problem, das auch Anwendungen bei der Berechnung von Flüssen mit minimalen Kosten in Graphen hat. Im Vergleich zu den lange bekannten exakten Algorithmen, hat der Algorithmus eine Laufzeitschranke mit verbesserter Abhängigkeit von $n$ und ist damit eine Verbesserung für dichte Graphen (das sind in diesem Fall Graphen mit $m = \Theta(n^2)$). Der Algorithmus berechnet für positive Kantengewichte eine multiplikative $(1+\varepsilon)$-Approximation des kleinesten durchschnittlichen Kantengewichts eines Kreises in Zeit $O(n^\omega \log^3(nW/\varepsilon)/\varepsilon)$, wobei $O(n^\omega)$ die beste asymptotische Laufzeit für die Multiplikation zweier $n \times n$ Matrizen ist. Wir reduzieren das Problem zuerst auf die wiederholte Anwendung von min-plus Matrixmultiplikation, die wiederum durch die Verwendung von klassischer Matrixmultiplikation approximiert werden kann. Dieser Ansatz liefert eine bessere asymptotische Laufzeit, die jedoch nicht praxisrelevant ist. Eine interessante zukünftige Forschungsfrage wäre daher die Entwicklung anderer Methoden um die Laufzeit für dieses grundlegende Graphproblem zu verbessern sowie die Erkundung der optimalen Balance zwischen Approximationsgarantie und Laufzeit.

**Algorithmen für MEPs mit Streett-Zielvorgaben.** Für Streett-Zielvorgaben zeigen wir für MEPs den ersten Algorithmus mit sub-quadratischer Laufzeit sowie einen Algorithmus mit verbesserter Laufzeit für dichte Graphen und MEPs. In ihrer vereinfachten Form sind die Laufzeiten $O(m^{1.5}\sqrt{\log n})$ und $O(n^2)$, was für MEPs die Laufzeit um einen Faktor von $n/\sqrt{\log n}$ bzw. $n$ verbessert und für Graphen die Laufzeit verbessert wenn $m \in \omega(n^{4/3}/\sqrt[3]{\log n})$. Während der einfachste Algorithmus für MEPs mit Streett-

Zielvorgaben bis zu $n$-mal eine sogenannte Zerlegung in maximale End-Komponenten für ein MEP berechnet, zeigen wir wie die Berechnung von maximalen End-Komponenten durch die einfachere Berechnung von starken Zusammenhangskomponenten ersetzt werden kann. Diese starken Zusammenhangskomponenten können dann wiederum mit graphalgorithmischen Techniken bei Veränderungen des MEPs schneller neu berechnet werden. Dafür verwenden wir eine Sparsifikationstechnik für dichte Graphen um die Laufzeit von $O(n^2)$ zu erhalten sowie einen lokalen Graphexplorationsansatz für die Laufzeit von $O(m^{1.5}\sqrt{\log n})$. Eine offene Fragestellung ist ob die Laufzeit weiter verbessert werden kann oder ob es vielleicht bedingte untere Schranken gibt. Erkenntnisse in jede dieser Richtungen könnten auch zu Fortschritten für andere Graphprobleme führen, für die ähnliche Techniken verwendet wurden.

**Paritätsspiele.** Für Paritätsspiele zeigen wir zuerst den ersten Algorithmus mit subkubischer Laufzeit von $O(n^{2.5})$ für Paritätsspiele mit drei Prioritäten, was eine Verbesserung der Laufzeit im Fall von $m \in \omega(n^{3/2})$ bedeutet. Zum Zeitpunkt der Dissertation verbesserte dieser Algorithmus die Laufzeit für alle Paritätsspiele mit einer konstanten Anzahl $c$ an Prioritäten, wurde inzwischen aber für $c > 3$ überholt [Fe17]. Während der klassische Algorithmus für Paritätsspiele mit drei Prioritäten wiederholt Büchispiele löst, zeigen wir dass abgeschlossene Teile der Gewinnmenge mit nur $\sqrt{n}$ Knoten bereits in einem Teilgraphen mit nur $O(n^{3/2})$ Kanten gefunden werden können und müssen daher nur für Teile der Gewinnmenge mit mehr als $\sqrt{n}$ Knoten den $O(n^2)$-Zeit Büchispiel-Algorithmus aufrufen.

Weiters zeigen wir einen *symbolischen* Algorithmus, der $O(n^{c/3+1})$ symbolische Schritte benötigt und eine lineare Anzahl von Mengen speichert. Durch eine Variation der Parameter liefert der gleiche Algorithmus auch die erste sub-exponentielle Schranke für die Anzahl der symbolischen Schritte. Dieser Algorithmus verbessert damit die Anzahl der benötigten symbolischen Schritte gegenüber dem bisher bestbekannten symbolischen Algorithmus, während er die gleiche Anzahl an Mengen speichert wie der grundlegende symbolische Algorithmus. Das Kernstück der neuen symbolischen Algorithmen ist eine symbolische Version eines "progress measure" genannten Zählers, der iterativ Daten über das Paritätsspiel sammelt, wobei für die symbolischen Variante $\Theta(n^{c/2})$ viele numerische Werte mit $O(n)$ vielen Mengen repräsentiert werden.

Das große offene Problem für Paritätsspiele ist die Existenz eines Polynomialzeitalgorithmus. Ein weiterer Weg Paritätsspiele besser zu verstehen könnten bedingte untere Schranken sein. Für die praktischen Anwendungen sind symbolische Algorithmen relevant und es wäre interessant ob die neuen Ideen für symbolische Berechnungen auch zu praktischen Verbesserungen führen.

**Modell- und Zielvorgaben-Separierung für Graphen und MEPs.** Wir zeigen mehrere neue Algorithmen und bedingte untere Schranken für Rabin-Zielvorgaben sowie Disjunktionen von Erreichbarkeits-, Sicherheits-, Büchi- und co-Büchi Zielvorgaben auf Graphen und MEPs. Diese Ergebnisse zeigen zum ersten Mal (1) eine *Separierung der Modelle* und (2) eine *Separierung der Zielvorgaben* für Polynomialzeitprobleme in formaler Verifikation. Für eine Separierung der Modelle, also in diesem Fall von MEPs und Graphen, zeigen

wir für das gleiche algorithmische Problem eine bedingte untere Schranke auf MEPs und eine strikt niedrigere obere Laufzeitschranke auf Graphen und zeigen damit, dass unter den Annahmen (A1) bzw. (A2) das algorithmische Problem auf MEPs schwieriger ist als auf Graphen. Für eine Separierung von Zielvorgaben vergleichen wir auf die gleiche Art und Weise zwei verwandte Zielvorgaben auf dem gleichen Modell. So zeigen wir insbesonders bedingte untere Schranke für Rabin-Zielvorgaben und strikt niedrigere obere Schranken für Streett-Zielvorgaben für sowohl Graphen als auch MEPs.

Die Basis für unsere unteren Schranken sind Reduktionen von CNF-SAT zu Disjunktionen von Erreichbarkeits- und Sicherheits-Zielvorgaben auf MEPs sowie von boolscher Matrixmultiplikation zu Disjunktionen von Sicherheits-Zielvorgaben auf Graphen und zu disjunktiven Erreichbarkeitsabfragen auf MEPs. Wir nützen dann Reduktionen zwischen den verschiedenen Zielvorgaben aus um auch untere Schranken für Büchi-, co-Büchi- und Rabin-Zielvorgaben zu erhalten.

Im Kern der neuen Algorithmen für Disjunktionen von Erreichbarkeits-, Büchi- und co-Büchi-Zielvorgaben auf MEPs stehen Beobachtungen zu den Eigenschaften der maximalen End-Komponenten, welche starke Zusammenhangskomponenten ohne ausgehende Zufallskanten sind. Weiters zeigen wir dass für die Disjunktion von co-Büchi-Zielvorgaben mit nur jeweils einem Knoten in der Zielmenge die Gewinnmenge auf Graphen mit einer Art von Breitensuche in linearer Zeit gelöst werden kann, was zu einer Modellseparierung für dieses Problem führt.

**Verallgemeinerte Büchi- und GR(1)-Spiele.** *Verallgemeinerte Büchi-Zielvorgaben* sind Konjunktionen von $k$ Büchi-Zielvorgaben und *GR(1)-Zielvorgaben* bestehen aus einer Implikation zwischen zwei verallgemeinerten Büchi-Zielvorgaben. Für verallgemeinerte Büchispiele verbessern wir die Laufzeit für dichte Graphen von $O(k^2 \cdot n^2)$ auf $O(k \cdot n^2)$ und zeigen dass diese Abhängigkeit der Laufzeit von $k$ und $n$ unter der Annahme (A1) optimal ist. Weiters zeigen wir dass der klassische Algorithmus für verallgemeinerte Büchispiele unter der Annahme (A2) eine optimale Abhängigkeit von $k$ und $m$ hat. Diese untere Schranke gilt selbst dann, wenn jede Büchi-Zielmenge nur einen Knoten enthält (in diesem Fall ist die Laufzeit $O(k \cdot m)$). Diese Ergebnisse implizieren weiters eine Modell-Separierung zwischen Spielgraphen einerseits und MEPs und Graphen anderseits. Weiters präsentieren wir einen Algorithmus für GR(1)-Spiele, der die Laufzeit für den Fall $m \in \omega(n^{1.5})$ verbessert.

## 4 Schlussworte

Diese Dissertation verbindet zwei Teilgebiete der theoretischen Informatik, zwischen denen es viel zu oft kaum Austausch gibt: Algorithmenentwicklung und Komplexitätstheorie auf der einen Seite und Modellprüfung, Automatentheorie und Spielgraphen auf der anderen Seite. In dem wir die asymptotischen Laufzeitschranken von verschiedenen Problemen wie generalisierten Büchispielen mit der von CNF-SAT und kombinatorischer boolescher Matrixmultiplikation verknüpfen, verbinden wir fundamentale algorithmische Probleme der beiden Gebiete der theoretischen Informatik und unsere Ergebnisse zeigen, dass algo-

rithmische Verbesserungen für fundamentale Probleme in formaler Verifikation und Synthese Durchbrüche in der Algorithmenentwicklung bedeuten würden. Weiters zeigen wir die Anwendbarkeit neuester Entwicklungen in der Algorithmenentwicklung und Komplexitätstheorie wie bedingte untere Schranken sowie Techniken aus dem Bereich der Graphalgorithmen für kanonische Probleme der Modellprüfung und Synthese. Aus der Sicht der Algorithmenforschung und Komplexitätstheorie ist unser Beitrag eine zugängliche Exposition wichtiger algorithmischer Probleme in der Sprache von Graphalgorithmen. insbesonders Paritäts- und Mittelwertspiele sind zwei der wenigen "natürlichen" Probleme in NP $\cap$ coNP für die noch kein Polynomialzeitalgorithmus bekannt ist und sie sind daher von größtem Interesse für die Algorithmenforschung und Komplexitätstheorie. Unsere Ergebnisse sind ein erster Schritt um die algorithmische Schwierigkeit von Polynomialzeitproblemen in formaler Verifikation und Synthese zu verstehen, es gibt weiterhin viele interessante offene Probleme, von denen einige in der Dissertation aufgelistet sind [Lo16].

## Literaturverzeichnis

[Br97] Browne, A.; Clarke, E. M.; Jha, S.; Long, D. E.; Marrero, W. R.: An Improved Algorithm for the Evaluation of Fixpoint Expressions. Theoretical Computer Science, 178(1-2):237–255, 1997.

[Br11] Brim, L.; Chaloupka, J.; Doyen, L.; Gentilini, R.; Raskin, J.-F.: Faster algorithms for mean-payoff games. FMSD, 38(2):97–118, 2011.

[BV07] Björklund, H.; Vorobyov, S. G.: A combinatorial strongly subexponential strategy improvement algorithm for mean payoff games. Discrete Applied Mathematics, 155(2):210–229, 2007.

[Ca17] Calude, C. S.; Jain, S.; Khoussainov, B.; Li, W.; Stephan, F.: Deciding Parity Games in Quasipolynomial Time. In: STOC. S. 252–263, 2017.

[CH14a] Chatterjee, K.; Henzinger, M.: Efficient and Dynamic Algorithms for Alternating Büchi Games and Maximal End-component Decomposition. Journal of the ACM, 61(3):15, 2014.

[Ch14b] Chatterjee, K.; Henzinger, M.; Krinninger, S.; Loitzenbauer, V.; Raskin, M. A.: Approximating the minimum cycle mean. Theoretical Computer Science, 547:104–116, 2014.

[Ch16a] Chatterjee, K.; Dvořák, W.; Henzinger, M.; Loitzenbauer, V.: Conditionally Optimal Algorithms for Generalized Büchi Games. In: MFCS. S. 25:1–25:15, 2016.

[Ch16b] Chatterjee, K.; Dvořák, W.; Henzinger, M.; Loitzenbauer, V.: Model and Objective Separation with Conditional Lower Bounds: Disjunction is Harder than Conjunction. In: LICS. S. 197–206, 2016.

[Ch17] Chatterjee, K.; Dvořák, W.; Henzinger, M.; Loitzenbauer, V.: Improved Set-Based Symbolic Algorithms for Parity Games. In: CSL. S. 18:1–18:21, 2017.

[CHL17] Chatterjee, K.; Henzinger, M.; Loitzenbauer, V.: Improved Algorithms for Parity and Streett objectives. Logical Methods in Computer Science, 13(3), 2017. Announced at LICS'15.

[CJH03] Chatterjee, K.; Jurdziski, M.; Henzinger, T. A.: Simple stochastic parity games. In: CSL. S. 100–113, 2003.

[EJ88] Emerson, E. A.; Jutla, C. S.: The Complexity of Tree Automata and Logics of Programs (Extended Abstract). In: FOCS. S. 328–337, 1988.

[EL86] Emerson, E. A.; Lei, Ch.-L.: Efficient Model Checking in Fragments of the Propositional Mu-Calculus. In: LICS. S. 267–278, 1986.

[Fe17] Fearnley, J.; Jain, S.; Schewe, S.; Stephan, F.; Wojtczak, D.: An Ordered Approach to Solving Parity Games in Quasi Polynomial Time and Quasi Linear Space. In: SPIN. S. 112–121, 2017.

[FV97] Filar, J.; Vrieze, K.: Competitive Markov Decision Processes. Springer-Verlag, 1997.

[HT96] Henzinger, M.; Telle, J. A.: Faster Algorithms for the Nonemptiness of Streett Automata and for Communication Protocol Pruning. In: SWAT. S. 16–27, 1996.

[JPZ08] Jurdziński, M.; Paterson, M.; Zwick, U.: A Deterministic Subexponential Algorithm for Solving Parity Games. SIAM J. Comput., 38(4):1519–1532, 2008.

[Ju00] Jurdziński, M.: Small Progress Measures for Solving Parity Games. In: STACS. S. 290–301, 2000.

[Ka78] Karp, R. M.: A characterization of the minimum cycle mean in a digraph. Discrete Mathematics, 23(3):309–311, 1978.

[Lo16] Loitzenbauer, V.: Improved Algorithms and Conditional Lower Bounds for Problems in Formal Verification and Reactive Synthesis. Dissertation, University of Vienna, 2016.

[OA92] Orlin, J. B.; Ahuja, R. K.: New scaling algorithms for the assignment and minimum mean cycle problems. Mathematical Programming, 54(1-3):41–56, 1992.

[Sc07] Schewe, S.: Solving Parity Games in Big Steps. In: FSTTCS. S. 449–460, 2007.

[Zi98] Zielonka, W.: Infinite games on finitely coloured graphs with applications to automata on infinite trees. Theoretical Computer Science, 200(1–2):135–183, 1998.

**Veronika Loitzenbauer**, geboren 1988, hat nach dem Bachelorstudium Computational Science an der Karl-Franzens Universität Graz und dem Masterstudium Scientific Computing an der Universität Wien ihre Leidenschaft für theoretische Informatik und insbesondere Graphalgorithmen entdeckt. Sie hat von 2012 bis 2017 an der Universität Wien in der Forschungsgruppe „Theorie und Anwendung von Algorithmen“, betreut von Prof. Monika Henzinger, promoviert. In dieser Zeit hat sie sowohl an algorithmischer Spieltheorie, fundamentalen Graphproblemen, als auch an denen in dieser Dissertation präsentierten algorithmischen Problemen in formaler Verifikation und Synthese geforscht. Ihre Dissertation wurde mit dem österreichischen Staatspreis „Award of Excellence“ ausgezeichnet. Sie hat Forschungsaufenthalte an der Università di Roma „Tor Vergata“, Italien, der University of Michigan, USA, und der Bar-Ilan Universität, Israel, absolviert. Seit Dezember 2017 forscht sie als PostDoc an der Johannes Kepler Universität Linz am Institut für formale Modelle und Verifikation.

# Verbesserung des Durchsatzes und der Zuverlässigkeit von drahtlosen Ultrahochgeschwindigkeitskommunikationen auf data link layer Ebene [1]

Lukasz Lopacinski[2]

**Abstract:** Das Entwerfen von drahtlosen 100 Gbps Netzwerken ist eine herausfordernde Aufgabe. Ein serieller Reed-Solomon-Decodierer für die angestrebte Datenrate muss mit einer ultra hohen Taktfrequenz von 12,5 GHz arbeiten, um die Zeitbegrenzungen der Übertragung zu erfüllen [Lo15]. Das Empfangen eines einzelnen Ethernet Frames auf der physischen Ebene kann schneller ablaufen, als der Zugriff auf den DDR3 Speicher [He09]. Darüber hinaus muss der Data-Link-Layer der drahtlosen Systeme mit einer hohen Bitfehlerrate (BER) arbeiten. Die BER in der drahtlosen Kommunikation kann um mehrere Größenordnungen höher liegen, als in drahtgebundener Kommunikation. Um Forward-Error-Correction auf aktuellsten FPGA zu betreiben, benötigt man einen höchst parallelisierten Ansatz. Daher müssen neue Verarbeitungskonzepte für schnelle drahtlose Kommunikation entwickelt werden. Aufgrund dieser genannten Fakten, und da er auch nicht von anderen Systemen übernommen werden kann, sollte der Data-Link-Layer für die drahtloses 100G Kommunikation als neue Forschung in Betracht gezogen werden. Diese Dissertation liefert eine detaillierte Fallstudie über ein 100 Gbps Data-Link-Layer Design, wobei der Hauptfokus auf der Verbesserung der Zuverlässigkeit für drahtlose Ultra-Hochgeschwindigkeits-Kommunikation liegt. Zuerst werden die Beschränkungen der verfügbaren Hardware-Plattformen identifiziert (Speicherkapazität, Speicherzugriffszeit und die Anzahl logischer Zellen). Danach wird ein FPGA Beschleuniger gezeigt, welcher auf dem Data-Link-Layer 118 Gbps an Benutzerdaten verarbeitet. Am Ende wird die ASIC-Synthese betrachtet und eine detaillierte Statistik der verbrauchten Energie gezeigt.

## 1 Einführung

Die Anzahl von drahtlos gesteuerten Geräten steigt mit jedem Jahr und fast in jedem Haus befinden sich Geräte, wie zum Beispiel Garagenöffner oder Smartphones. Normalerweise brauchen sie keine extrem hohen Datenraten. Allerdings findet man schon heute Anwendungen, für die verfügbare Technologien zu langsam sind. Eine solche Anwendung basiert auf der sog. Virtual-Reality (VR) Technologie. Eines der größten Probleme beim VR-Anwendungen ist die begrenzte Leistung aktueller drahtloser Kommunikationstechnologien. Zum Beispiel verlangen manche VR-Szenarien extrem hohe Datenraten, und sehr kurze Latenzzeiten von $<$1ms [FA14]. Zusätzlich müssen die neuen drahtlosen Technologien in kleine, batteriebetriebenen Geräten integriert werden. Solche extremen Anforderungen machen die Forschung für künftige drahtlose Kommunikation besonders anspruchsvoll: Drahtlose Technologie für Zukunftsanwendungen muss um den Faktor 100 beschleunigt werden, der Energieverbrauch darf aber nicht steigen. Deswegen muss der

[1] Englischer Titel der Dissertation: "Improving goodput and reliability of ultra-high-speed wireless communication at data link layer level"

[2] IHP, lopacinski@ihp-microelectronics.com

Energieverbrauch pro Bit etwa um das Hundertfache gesenkt werden. Diese Arbeit befasst sich mit künftiger drahtloser Kommunikation, mit extrem hohen Datenraten von 100 Gbps und mehr und untersucht insbesondere die Sicherungsschicht (Schicht 2 des OSI-Modells) für solch schnelle, drahtlose Systeme. Zusätzlich zur Realisierung dieser extrem hohen Datenrate soll das gesamte Übertragungs-system nur ca. 1 Watt Leistung verbrauchen. Aus diesem Grund müssen alle verwendeten Algorithmen entsprechend der verbrauchten Energie signifikant verbessert und vollständig parallel ausgeführt werden. Daher schlägt die Arbeit einen neuen Dekodierer-Algorithmus für 100-Gbps-Turbo-Product-Codes [Tz16] vor, der effizienter gegenüber Bitfehlern ist und ca. 25% weniger Energie verbraucht. Zusätzlich wird eine weitere Reed-Solomon-Lösung [Wa16] für die gegebene Anwendung übernommen und bietet eine extrem hohe Datenrate, 169 Gbps @ 220 MHz im Virtex7 FPGA, bei sehr geringem Hardware- und Energieaufwand (mehr dazu in Abschnitt 4). Alle Paketbestätigungs-, Fragmentierungs-, Aggregierungs-, Linkadaptierungs- und Frame-Verarbeitungs-Algorithmen werden ohne Datenabhängigkeiten implementiert und vollständig parallel ausgeführt mit einem Netto-Durchsatz von ca. 118 Gbps. Die Konzepte und Modelle dieser Arbeit wurden nicht nur mit analytischer Evaluierung bestätigt, sondern auch implementiert und mit FPGA-Platt-formen gründlich evaluiert. Dieser Ansatz erreicht eine Datenrate von ca. 118 Gbps, die etwa 400x schneller als der neuste Mobilfunkstandard LTE ist. Die auf 40 nm basierte ASIC Implementierung verbraucht dabei nur ca. 10 pJ/Bit. Die hier dargestellte Lösung für drahtlose Verbindungen mit Datenraten über 100 Gbps ist eine der Ersten (wahrscheinlich die Erste) weltweit.

## 2 Herausforderungen der drahtlosen 100 Gbps Data Link Layer

Obwohl heutige drahtgebundene Kommunikationsstandards, wie Ethernet oder Glasfaser, hohe Datenraten von 100 Gbps erreichen, lassen sich solche Ansätze für drahtlose Kommunikation nicht einsetzen. Der Grund dafür liegt an der viel höheren Fehlerrate bei drahtlosen Verbindungen. Deswegen können die Implementierungen in den drahtgebundenen Netzwerken nicht effizient mit hohen Bitfehlerraten umgehen. Eine andere Besonderheit der drahtlosen Kommunikation ist die Komplexität im Duplexbetrieb. In drahtgebundenen Netzen braucht man einfach ein separates Kabel für jede Kommunikationsrichtung und die Netzgeräte können gleichzeitig senden und empfangen. Im Gegensatz dazu erlaubt eine typische drahtlose Verbindung, die Daten nur zu senden oder zu empfangen, und arbeitet typischerweise im sog. half-duplex. Aus diesen Gründen lassen sich die aktuellen Lösungen für drahtgebundene Netze nicht für drahtlose Kommunikation einsetzen. Daher müssen neue innovative Lösungen erforscht werden.

Die Sicherungsschicht braucht Speicher für die empfangenen und ausgehenden Pakete. Selbstverständlich darf auch der Speicher keine großen Latenzzeiten verursachen, damit hohe Datenraten unterstützt werden. Zusätzlich braucht diese Schicht mindestens 12 GB Speicherplatz für die Daten aus der letzten Sekunde. So viel Platz kann man nur in einen dezidierten Speicher schreiben, zum Beispiel DDR3. Die DDR3-Speicherzugriffslatenz beträgt aber ca. 45 Nanosekunden [He09], ist also viel zu hoch. Aus diesem Grund, man muss ein dezidierter Cache-Speicher genutzt werden und zusätzlich der sog. Zero-Copy-Ansatz angepasst und integriert werden.

Die Vorwärtsfehlerkorrektur verbessert die Robustheit gegen Übertragungsfehler, was besonders wichtig in drahtloser Kommunikation ist. Solche Fehlerkorrekturen verlangen aber viel Ressourcen, Chipfläche und auch Bearbeitungszeit. Zum Beispiel benötigt ein Viterbi-Dekodierer (1/2 Rate, mit 5-Bit Softcoding) eine Chipfläche so groß wie ca. 20 $mm^2$ im 40 nm CMOS [Ma10], um den Datendurchsatz von 100 Gbps zu erreichen. Ein anderer FEC-Ansatz - Reed-Solomon(255,239) für 100 Gbps Datenrate - leidet unter hohem Bearbeitungsaufwand mit etwa 70 000 000 MIPS . Aus diesem Grund stellt die 100 Gbps drahtlose Kommunikation sehr hohe Anforderungen an die Implementierung der Vorwärtsfehlerkorrektur, die zum einen eine große Robustheit gegen Übertragungsfehler aufweisen und zum anderen effizient und mit weniger Ressourcen arbeiten soll.

Da diese Arbeit die drahtlose Kommunikation für mobile Geräte berücksichtigt, spielt der Energieverbrauch eine große Rolle. Die Paketbearbeitung von hohen Datenraten führt aber zu hohem Energieverbrauch. Zum Beispiel eine Softwareimplementierung der Sicherungsschicht für 100 Gbps Glasfaserleitung, die auf Intel-Xeon-Prozessoren basiert, verbraucht etwa 650 Watt [He09]. Da die drahtlose Kommunikation viel komplexere Ansätze verlangt, würde der Energieverbrauch sogar höher sein. Allerdings sollte das Übertragungssystem für mobile Geräte nicht mehr als 1 Watt verbrauchen. Für die Sicherungsschicht, die ein Teil des Kommunikationssystems ist, bleibt sogar weniger als 1 Watt übrig.

## 3 Verfahren und Methoden

Um sehr hohe Datenraten bei drahtloser Kommunikation zu erreichen, müssen viele Ansätze eingesetzt und angepasst werden. Diese Arbeit untersucht und vergleicht verschiedene Lösungsansätze, prüft die Anpassung deren Parameter an die Performance in drahtloser Umgebung und führt neue Konzepte zur Lösung ein. Arbeitsmethodisch wurden zuerst die unterschiedlichen Ansätze mittels MATLAB-Simulationen untersucht. Anschließend erfolgten Testen und Bewerten mit der Hardwareimplementierung auf FPGA-Boards. Zu den wichtigsten Lösungen der Sicherungsschicht für schnelle und robuste drahtlose Kommunikation gehören: Fragmentierung und Aggregation von Paketen, Vorwärtsfehlerkorrekturmechanismen, Paketbestätigung und -wiederholung (engl. Automatic Repeat Request). Im Folgenden werden diese Lösungen näher betrachtet.

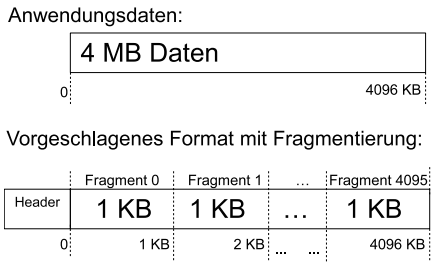



Abb. 1: Vorgeschlagenes Frameformat.

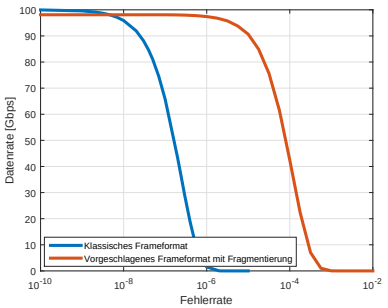



Abb. 2: Vorgeschlagenes Frameformat.

### 3.1 Frameformat, Fragmentierung, Aggregation und parallele Framebearbeitung

Die Framegröße und ihr Format beeinflussen stark die Kommunikationsperformance. Als Eingangspunkt wurde das Frameformat des IEEE-802.11ad-Standards ausgewählt (WLAN mit 60-GHz-Band [IEE12]). Nach vielen MATLAB-Simulationen wurde das Frameformat auf die betrachteten Szenarien angepasst, d. h. für drahtlose Netze mit einem Durchsatz von 100 Gbps und für unterschiedliche Bitfehlerraten. Basierend auf analytischen Untersuchungen wurde das Frameformat für 100-Gbps-Szenarien definiert (Abb. 1). Das Frame soll mindestens 4 MB groß sein, geteilt in 4096 Fragmente. Dank der Fragmentierung müssen bei Fehlern nur einzelne Teile und nicht ganze Frames wiederholt werden. Mit einem solchen Ansatz steigt die Robustheit gegenüber Bitfehlern um den Faktor 100 (Abb. 2).

Die kurze Bearbeitungszeit von Paketen erreicht man nur mit Hardwareimplementierungen, die alle Operationen parallel ausführen. In diesem Ansatz wird jeder Datenfluss in sog. Lanes aufgeteilt, danach separat parallel bearbeitet und schließlich zusammengefasst. Die Frameaggregation und das ARQ-Protokoll lassen sich gut in eine solche parallele Verarbeitung umsetzen. Viel schwerer ist die Realisierung einer parallelen Implementierung der Vorwärtsfehlerkorrektur, weil alle separaten Lanes voneinander komplett unabhängig sein müssen. Aus diesem Grund wurden in dieser Arbeit zwei FEC-Ansätze näher betrachtet ‘Interleaved Reed-Solomon Codes (IRS)‘ [Wa16] und ‘Turbo Product Codes (TPC)‘ [Tz16].

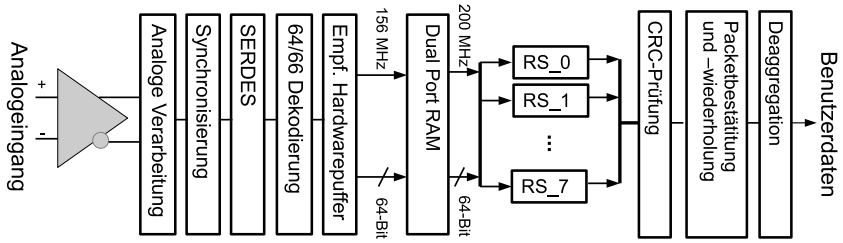



Abb. 3: Empfänger-Lane-Realisierung optimiert für IRS-Dekodierung und Virtex7-FPGA.

### 3.2 Interleaved Reed-Solomon

Nach dem Simulieren unterschiedlicher Vorwärtsfehlerkorrekturalgorithmen in MATLAB und Schätzung ihres Bedarfs an Chipfläche wurde der Reed-Solomon-Ansatz für die 100-Gbps-Sicherungsschicht ausgewählt. Danach wurde dieser Ansatz zusätzlich noch mit einem Interleaver verbessert, der die Korrektur von Burstfehler noch effektiver werden lässt. Abb. 3 stellt den Ansatz dar. Da diese Idee eine viel kleinere Komplexität als Konvolutionskodes und auf Interleavermatrizen basierte Lösungen aufweist, eignet sie sich besonders gut für die 100-Gbps-Sicherungsschicht. Abb. 3 zeigt eine FPGA- und ASIC-Implementierung einer Lane der Sicherungsschicht, die einen Durchsatz von 10 Gbps erreicht. Um die Datenrate von 100 Gbps zu unterstützen, müssen 10 solche Lanes parallel

gekoppelt werden. Die auf RS basierte FPGA und ASIC Implementierung erreicht einen Durchsatz von ca. 118 Gbps, und die 40 nm ASIC Lösung braucht nur etwa 0.8 $mm^2$ der Chipfläche.

### 3.3 Turbo Product Codes

Die auf Turbo Product Codes (TPC) [Tz16] basierte Vorwärtsfehlerkorrektur wird in 100-Gbps-Glasfaserkommunikation eingesetzt. Da die drahtlose Kommunikation mit höheren Fehlerraten umgehen muss, verlangen TPC-Lösungen Anpassungen. Im Rahmen dieser Arbeit wurden TPC-Ansätze untersucht und viele Konzeptverbesserungen für den Einsatz mit drahtloser Kommunikation umgesetzt, was in den nächsten Absätzen erklärt wird.

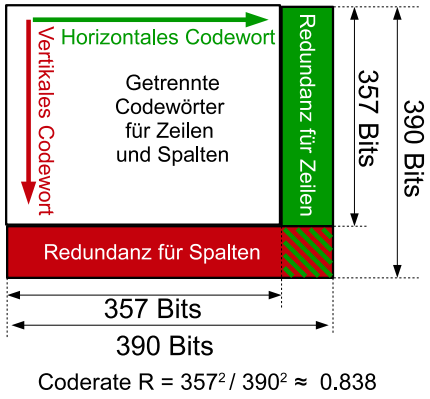



Abb. 4: Typischer Turbo-Product-Code-Dekodierer.

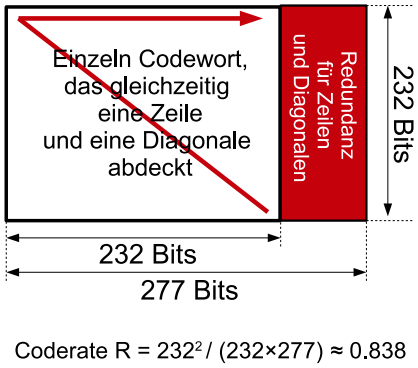



Abb. 5: Vorgeschlagener Turbo-Product-Code-Dekodierer.

#### 3.3.1 Änderungen im Codewort

Ein TPC-Dekodierer basiert auf iterativer Dekodierung von Bits, welche in einer Matrix organisiert sind (Abb. 4). Der Dekodierer bearbeitet einzelne Zeilen und Spalten aus der Matrix nacheinander und versucht sie zu dekodieren. Das Hauptproblem, wodurch sich die Fehlerkorrekturleistung verringert, sind die geteilten Komponenten-Codewörter für die Zeilen und Spalten. Ein gemeinsamer Dekodierer, der die Zeilen und Spalten gleichzeitig abdeckt, verwendet die Redundanz-Bits effizienter, da die Spalten- und Zeilenredundanz-Bits gleichzeitig verwendet werden. Außerdem ermöglicht diese Modifikation eine höhere Hammingdistanz zwischen den Komponenten-Codewörtern mit gleicher Coderate. Dadurch korrigiert die neue Methode mehr Bitfehler. Diese Arbeit verwendet zusätzlich noch eine diagonale Form des Codeworts anstelle einer Zeile-Spalten-Form (Abb. 5). Dadurch, dass die Datenbits mit Spaltenreihenfolge geschrieben werden, werden Burstfehler in der diagonalen Form des Codeworts vermieden. Die letzte vorgeschlagene Änderung ist noch, den BCH-Code für TPC: BCH(511,466,t=5) anstelle des BCH(390,357,t=3) zu nutzen. Zwar steigt der momentane Leistungsverbrauch um 30 %, aber die Dekodierungszeit ist

um ca. 40 - 50 % reduziert. Relativ gesehen korrigiert dieser Ansatz gleiche Fehlerraten mit weniger Energieverbrauch.

### 3.3.2 Hardware Turbo Product Code Dekodierer mit entrollenden Iterationen

Obwohl die erwähnten TPC-Anpassungen die Performance verbessern und den Energieverbrauch reduzieren, lassen sie sich in eine parallele Implementierung nicht umsetzen. Wegen Dekodierungsabhängigkeiten müssen die Frames nacheinander, nicht parallel, bearbeitet werden. Damit keine Dekodierungsabhängigkeiten entstehen also um die Frames parallel zu bearbeiten, werden zur Dekodierung zwei unabhängige Codes für ungerade und gerade Iterationen verwendet. Die Abb. 6 zeigt diese Codes als grüne und rote Markierungen. Die effektive Coderate (CR) kann durch Modifizieren der horizontalen Länge der Decodier-Matrizen geändert werden. Abb. 7 zeigt die vorgeschlagene Hardware-Implementierung mit entrollenden Iterationen.

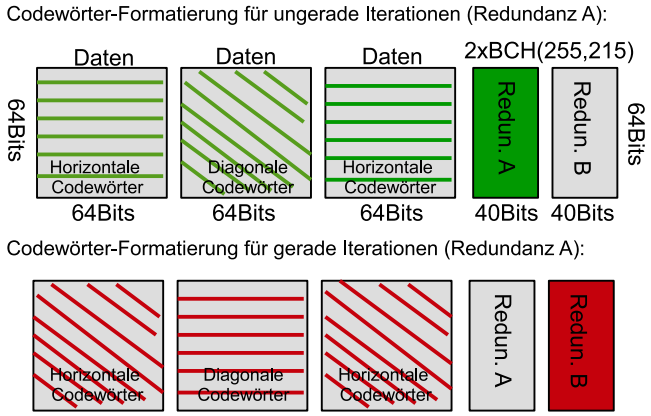



Abb. 6: Modifizierte TPC-Dekodierer für FPGA- und ASIC-Implementierungen. Hier werden zwei unabhängige Codes verwendet (als grün und rot dargestellt), um die Datenabhängigkeit zu lösen. Damit können Frames parallel bearbeitet werden.

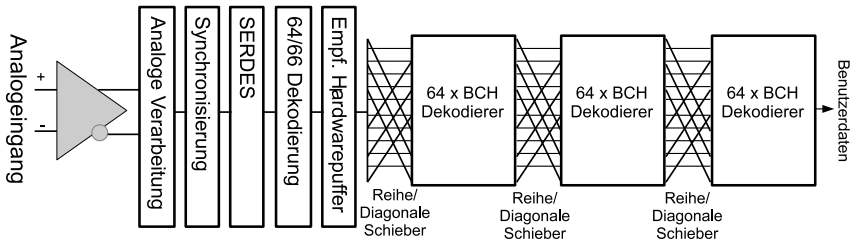



Abb. 7: Empfänger-Lane-Realisierung optimiert für TPC-Dekodierung und Virtex7-FPGA.

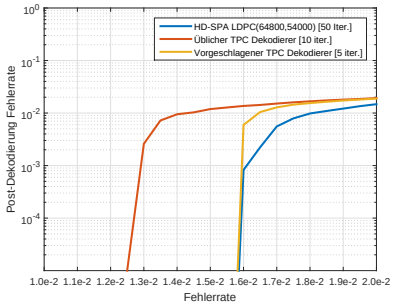



Abb. 8: Fehlerkorrekturleistung für den üblichen TPC (rot), vorgeschlagenen TPC (gelb) und HD-SPA LDPC (blau) Methoden.

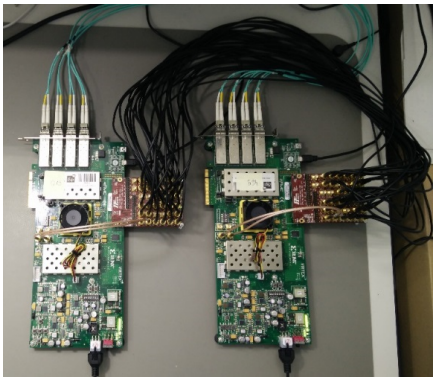


Abb. 9: Der FPGA-Demonstrator.

## 4 Ergebnisse

Diese Arbeit vergleicht den verbesserten TPC-Dekodierer mit anderen TPC- und LDPC-Lösungen. Zum Beispiel erreicht in dieser Arbeit die angepasste TPC-Implementierung mit einer Coderate (CR) von 0,8 ähnliche Ergebnisse wie der LDPC, aber viel bessere Ergebnisse als der übliche TPC-Ansatz (Abb. 8). Im AWGN-Kanal korrigiert der neue TPC um 28 % höhere Fehlerraten als der übliche TPC-Dekodierer. Ein weiterer Vorteil ist der geringere Energieverbrauch. Obwohl die Codier-Leistung um ca. 30% gestiegen ist, braucht der Dekodierer weniger Iterationen und damit 20-25% weniger Energie.

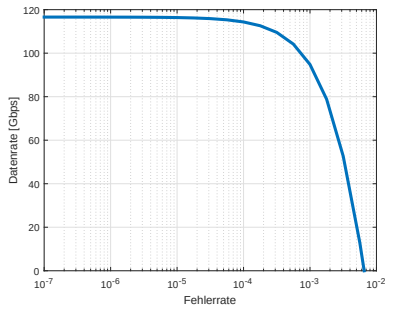



Abb. 10: Nutzdatendurchsatz als Funktion der Fehlerrate für den FPGA-Demonstrator.

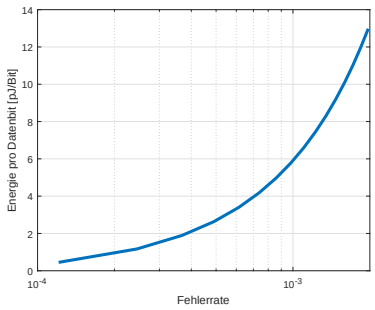



Abb. 11: Energieaufnahme pro Datenbit mit aktivierter Linkadaptierung (40-nm-CMOS).

Die gute Performance der in dieser Arbeit untersuchten und implementierten Lösungen wurde mithilfe der FPGA-Plattform bestätigt (Abb. 9). Die Implementierung umfasst Frame-Aggregierung und -Fragmentierung, Interleaved Reed-Solomon Fehlerkorrektur, CRC-modul, Linkadaptierung und hybrid ARQ. Diese Implementierung hat den Datendurchsatz von ca. 118 Gbps erreicht und damit ist sie eine der schnellsten Sicherungsschichten für drahtlose Systeme weltweit (Abb. 10).

Die VHDL-Implementierung der Data Link Layer, die ursprünglich auf FPGA portiert und getestet wurde, erreicht auch gute Ergebnisse in ASIC-Simulationen. Dank der effi-

zienten und parallelen Implementierung erreicht die maximale Taktfrequenz 210 MHz für die IHP-130-nm- und 900 MHz für die 40-nm-ASIC-Technologie. Die auf 40 nm basierte Technologie erreicht einen Datendurchsatz von 55,87 Gbps für ein Lane. Damit kann man mit nur 2 Lanes die Datenrate von mehr als 100 Gbps unterstützen.

Da die Leistungsaufnahme in 40 nm sehr gering ist, weniger als 1,3 W, kann es in akkubetriebenen mobilen Geräten (z.B. Laptops) eingebaut werden. Den Energieverbrauch für ein einzelnes Bit stellt die Abb. 11 dar. Für Bitfehlerraten weniger als 2e-4 verbraucht der Prozessor nur 1 pJ/Bit. Für Bedingungen mit häufigen Fehlern, mit einer Fehlerrate von 2e-3, steigt der Energieverbrauch wegen häufigerem FEC-Einsatz auf 13 pJ/Bit.

Die hier untersuchte Sicherungsschicht soll mit der Modulation PSSS (engl. Parallel Sequence Spread Spectrum) und PAM (engl. Pulse-Amplitude Modulation) zusammenarbeiten, mit einer spektralen Effizienz von 4 Bit/s/Hz. Ohne Verstärkung in PSSS braucht die in der Arbeit implementierte Sicherungsschicht 19,5 dB Eb/N0 und 6,5 pJ/Bit, um den Durchsatz von 100 Gbps zu erreichen.

LDPC-Dekodierer sind derzeit die beliebtesten FEC-Methoden, die in industriellen Anwendungen verwendet werden. Tabelle 1 vergleicht die Data-Link-Layer-Implementierung mit LDPC-Methoden, die in 40-nm-CMOS synthetisiert sind. Die in dieser Arbeit dargestellte Lösung verbraucht deutlich weniger Ressourcen und ermöglicht eine viel schnellere Verarbeitung von Daten. Sie kann dafür aber weniger Fehler korrigieren und benötigt höhere Eb/N0.

| | LDPC [Li13] | LDPC [Mo15] | Diese Arbeit |
|---|---|---|---|
| Technologie | 40 nm G | 40 nm LP | 40 nm |
| EFC-Algorithmus | Soft decision LDPC 802.11ad | Soft decision LDPC 802.11ad | Hard decision IRS |
| Eb/N0 bei Post-FEC Fehlerrate = 1e-5 | 5,9 dB; CR=13/16 QPSK, Indoor-Mehrwege-Kanal | 3,5 dB; CR=1/2, QPSK | 6,2 dB; CR=223/255 QPSK, AWGN, worst case |
| Netto Durchsatz | 5,6 Gbps | 6,16 Gbps | 111,74 Gbps |
| Chipfläche | 0,16 $mm^2$ | 0,8 $mm^2$ | 0,81 $mm^2$ |
| Energieaufnahme | 18 pJ/Bit | 32,9 pJ/Bit | Min. 1 pJ/Bit (BER $<$ 2e-4); Max. 13 pJ/Bit (BER $>$ 2e-3) |
| Normalisierte Chipfläche | 35 Gbps/$mm^2$ | 7,7 Gbps/$mm^2$ | 140 Gbps/$mm^2$ |
| Funktionalität | FEC-Dekodierer | FEC-Dekodierer | FEC-Enkodierer + FEC-Dekodierer + Sicherungsschichtprozessor |

Tab. 1: Vergleich des implementierten Prozessors mit zwei modernen 802.11ad-LDPC- Dekodierern (40-nm-CMOS-Technologien).

## 5 Zusammenfassung

Der Schwerpunkt dieser Arbeit lag in der Untersuchung und auch Umsetzung der Sicherungsschicht (Schicht 2 des OSI-Modell) für künftige ultraschnelle drahtlose Datenkommunikation mit Datenraten über 100 Gbps. So hohe Datenraten für drahtlose Verbindungen wurden bisher nicht berücksichtigt und stellen eine große Herausforderung dar, insbesondere auch an die Sicherungsschicht. Die bereits bekannten Ansätze für drahtlose Kommunikation sind viel zu langsam und verbrauchen zu viel Energie. Demzufolge hat diese Arbeit eine neue Sicherungsschicht untersucht und erfunden, welche Datenraten von ca. 118 Gbps erreicht, den Energieverbraucht stark reduziert und als ASIC realisierbar ist. Im Folgenden werden die wichtigsten Ergebnisse dieser Arbeit kurz zusammengefasst:

1. Dank des innovativen, in dieser Arbeit erfundenen Dekodierers eignet sich die Turbo-Product-Codes (TPC) Vorwärtsfehlerkorrektur gut für ASIC Implementierungen und verbraucht 25% weniger Energie gegenüber üblichen TPC Ansätzen.
2. Die auf Reed-Solomon (RS) basierte Vorwärtsfehlerkorrektur wurde gründlich mit MATLAB untersucht und schließlich mit einem Interleaver verbessert. Die angepasste FPGA Implementierung erreicht eine Datenrate von 169 Gbps (im Virtex7 FPGA mit 220 MHz Clock). Zusätzlich verbraucht die ASIC Lösung sehr wenig Energie, ca. 10 pJ/Bit in 40 nm CMOS Technologien. Die verbesserte RS Fehlerkorrektur korrigiert bis zu 1024 Bits lange Burst-Fehler.
3. Die komplette Sicherungsschicht wurde in Form eines Data-Link-Layer-Prozessors in der FPGA-Plattform implementiert und erreicht den Nettodurchsatz von ca. 118 Gbps. Es scheint der weltweit schnellste Data-Link-Layer-Prozessor für drahtlose Anwendungen zu sein, da bisher keine Veröffentlichungen über so schnelle Sicherungsschichten berichten.
4. Diese Arbeit hat auch unterschiedliche Hardware-Plattformen für die Realisierung der neuen Sicherungsschicht näher betrachtet: Xilinx Virtex7 FPGA, IHP-130-nm-CMOS-Technologie und die industrielle 40-nm-CMOS-Technologie. Für alle drei Fälle wurden die hardware-spezifischen Optimierungsparameter betrachtet und die Datenraten von $>$100 Gbps erreicht.

Diese Arbeit erreicht sehr gute Ergebnisse nicht nur in Datenraten sondern auch im kleinen Energieverbrauch. Zum Beispiel lassen sich die hier dargestellten Lösungen sofort in Laptops integrieren, da der Energieverbrauch nur ca. 10 pJ/Bit beträgt. Für kleinere mobile Geräten (z. B. Smartphones) muss aber die Energie auf 1-2 pJ/Bit reduziert werden. Dazu muss man aber den sog. vertikalen Ansatz berücksichtigen, und auch Optimierungen in der physikalischen Ebene (auch in Nanostruktur) einbeziehen. Dies geht über die Grenzen dieser Arbeit hinaus. Die weiteren wissenschaftlichen Arbeiten in der Sicherungsschicht für künftige, schnelle drahtlose Kommunikation liegen hauptsächlich in der Vorwärtsfehlerkorrektur, die etwa 95% der gesamten Energie der Sicherungsschicht verbraucht. Auf der einen Seite soll die Forschung in die Richtung noch energiesparsamerer Fehlerkorrekturalgorithmen gehen. Auf der anderen Seite müssen auch höhere Kommunikationsschichten und insbesondere auch die spezielle Anwendungsschicht berücksichtigt

werden, also der sog. vertikale Ansatz. Zum Beispiel können manche Anwendungen und Protokolle höhere Bitfehlerraten akzeptieren und damit eine einfachere und energiesparsamere Fehlerkorrektur ermöglichen.

## Literaturverzeichnis

[FA14] Fettweis, Gerhard; Alamouti, Siavash: 5G: Personal mobile internet beyond what cellular did to telephony. IEEE Communications Magazine, 52(2):140–145, 2014.

[He09] Hermsmeyer, Christian; Song, Haoyu; Schlenk, Ralph; Gemelli, Riccardo; Bunse, Stephan: Towards 100G packet processing: Challenges and technologies. Bell Labs Technical Journal, 14(2):57–79, 2009.

[IEE12] IEEE 802.11ad-2012 Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3 : Enhancements for Very High Throughput in the 60 GHz Band, 2012.

[Li13] Li, Meng; Naessens, Frederik; Debacker, Peter; Raghavan, Praveen; Desset, Claude; Li, Min; Dejonghe, Antoine; Van der Perre, Liesbet: An area and energy efficient half-row-paralleled layer LDPC decoder for the 802.11 AD standard. In: Signal Processing Systems (SiPS), 2013 IEEE Workshop on. IEEE, S. 112–117, 2013.

[Lo15] Lopacinski, L.; Nolte, J.; Buechner, S.; Brzozowski, M.; Kraemer, R.: 100 Gbps wireless - data link layer VHDL implementation. Measurement Automation Monitoring, Vol. 61, No. 7:333–336, 2015.

[Ma10] Marinkovic, Miroslav; Piz, Maxim; Choi, Chang-Soon; Panic, Goran; Ehrig, Marcus; Grass, Eckhard: Performance evaluation of channel coding for Gbps 60-GHz OFDM-based wireless communications. In: Personal Indoor and Mobile Radio Communications (PIMRC), 2010 IEEE 21st International Symposium on. IEEE, S. 994–998, 2010.

[Mo15] Motozuka, Hiroyuki; Yosoku, Naoya; Sakamoto, Takenori; Tsukizawa, Takayuki; Shirakata, Naganori; Takinami, Koji: A 6.16 Gb/s 4.7 pJ/bit/iteration LDPC decoder for IEEE 802.11 ad standard in 40nm LP-CMOS. In: Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on. IEEE, S. 1289–1292, 2015.

[Tz16] Tzimpragos, Georgios; Kachris, Christoforos; Djordjevic, Ivan B; Cvijetic, Milorad; Soudris, Dimitrios; Tomkos, Ioannis: A survey on FEC codes for 100 G and beyond optical networks. IEEE Communications Surveys & Tutorials, 18(1):209–221, 2016.

[Wa16] Wang, Zhongfeng; Chini, Ahmad; Kilani, Mehdi T; Zhou, Jun: Multiple-symbol interleaved RS codes and two-pass decoding algorithm. China Communications, 13(4):14–19, 2016.

**Lukasz Lopacinski** absolvierte 2009 seinen Master an der Westpommerschen Technischen Universität (Szczecin, Polen). Seit 2007 arbeitet er in der Industrie in den Bereichen eingebettete Systeme und drahtlose Kommunikation. 2017 promovierte er in diesem Bereich an der BTU Cottbus und arbeitet seit dem am IHP (Frankfurt-Oder, Deutschland).

# Selektives Lernen für Empfehlungsmaschinen[1]

Pawel Matuszyk[2]

**Abstract:**

Das Problem der Informationsüberladung ist im digitalen Zeitalter besonders herausfordernd. Um Nutzer beim Finden relevanter Artikel zu unterstützen, wurden Empfehlungsalgorithmen entwickelt. Diese Algorithmen lernen Nutzerpräferenzen und sagen voraus, was Nutzer in der Zukunft relevant finden werden. Empfehlungsmaschinen lernen aus historischen Daten zum Nutzerfeedback. Diese Daten sind jedoch von fehlenden Werten dominiert, da jeder Nutzer Feedback zu nur wenigen Produkten bereitstellen kann. Das hat zur Folge, dass konventionelle Ansätze zum Lernen von Nutzerpräferenzen *alle verfügbaren Daten* verwenden. In dieser Arbeit schlagen wir *selektives Lernen* vor. In diesem *alternativen Ansatz* werden *Daten, Nutzer, oder Aspekte der Modelle nur selektiv zum Trainieren der Präferenzmodelle* verwendet. Wir zeigen, dass dieser Ansatz zu einer *signifikanten Verbesserung der Empfehlungsqualität* führt.

## 1 Einführung

Empfehlungsmaschinen (recommender systems) mildern das Problem der Informationsüberladung, das immer dann auftritt, wenn Menschen aus einer Vielzahl von Alternativen wählen müssen, von denen nur wenige relevant sind. Heutzutage sind Menschen mit diesem Problem häufiger denn je konfrontiert (z.B. bei einer Vielzahl von Filmen, Büchern, Musikstücken oder Medikamenten). Die Anwendungen der Empfehlungsmaschinen umfassen jedoch nicht nur E-Commerce [Li14, SKR99, PNH15], sondern auch Medizin [WP14, Ch12], Lehrmaterialien [Ma11, WLZ15], u.v.a.m.

Empfehlungsmaschinen lernen Nutzerpräferenzen aus historischem Feedback und sagen voraus, welche Artikel für einen gegeben Nutzer in der Zukunft relevant sein werden. Die relevanten Artikel werden dann individuell einem gegebenen Empfänger empfohlen. Das historische Feedback von Nutzern ist allerdings oft sehr spärlich, da jeder Nutzer nur relativ wenige Artikel wahrnehmen kann. Dies führt zu einer häufigen Annahme in der Forschung zu Empfehlungsmaschinen, dass alle verfügbaren Daten zum Lernen der Präferenzmodelle genutzt werden sollen. In diesem Beitrag und in der zugrundeliegenden Dissertation schlagen wir ein neues Paradigma vor, laut dem die Trainingsdaten vorsichtig selektiert werden sollten. Wir bezeichnen das neue Paradigma als *selektives Lernen*. Wir entwickeln drei Typen von Ansätzen zum selektiven Lernen, sowohl für strombasierte, als auch für batch-basierte Algorithmen.

Die strom-basierten Algorithmen, die im Fokus dieser Arbeit liegen, haben gegenüber den batch-basierten Algorithmen einen entscheidenden Vorteil. Sie sind adaptiv und können

[1] Englischer Titel der Dissertation:"Selective Learning for Recommender Systems" [Ma17a]

[2] Fakultät für Informatik, Otto-von-Guericke-Universität Magdeburg, Deutschland, pawel.matuszyk@ovgu.de

neue Information sofort in ihre Präferenzmodelle integrieren. Somit können sie, idealerweise in Echtzeit, auf Änderungen reagieren.

Unser erster Ansatz zum selektiven Lernen sind Vergessensmethoden für strombasierte Empfehlungsmaschinen. Sie selektieren, welche Daten vergessen werden sollten. Somit entscheiden sie auch, welche Daten zum Lernen der Modelle verwendet werden. Die Vergessensmethoden stellen eine weitere Möglichkeit zur Adaptation der Modelle an Änderungen der Nutzerpräferenzen dar. Wir betonen, dass nicht nur veraltete Daten vergessen werden können. Wir schlagen elf Vergessensstrategien vor, die die zu vergessende Informationen selektieren, und drei alternative Algorithmen, die das Vergessen umsetzen.

Unser zweiter selektiver Ansatz eignet sich für Selektion der Nachbarn in Collaborative-Filtering-Algorithmen (CF). Die CF-Algorithmen arbeiten ähnlich wie k-Nearest-Neighbours im maschinellen Lernen. Sie suchen nach ähnlichen Nutzern und empfehlen die von ihnen als relevant empfundenen Produkte. Unser selektives Kriterium erlaubt die Selektion der Nachbarn nicht nur aufgrund der Ähnlichkeit, sondern auch aufgrund der Zuverlässigkeit der Information.

Unser letzter selektiver Ansatz basiert auf teil-überwachtem Lernen. Wir haben das erste teil-überwachte Framework für strom-basierte Empfehlungsmaschinen entwickelt. Dieses Frawemork erlaubt es den Empfehlungsmaschinen, von den umfangreichen fehlenden Daten zu lernen. Das ist z.B. im Co-Training-Ansatz möglich, wo mehrere Algorithmen parallel arbeiten und sich gegenseitig trainieren, indem sie eigene Vorhersagen anderen Algorithmen als Trainingsbeispiele zur Verfügung stellen. Nicht alle Vorhersagen allerdings vertrauenswürdig. Das Lernen aus falschen Vorhersagen könnte eine Verschlechterung der Empfehlungen zur Folge haben. Um das zu verhindern, schlagen wir weitere Selektionsmechanismen vor, die es erlauben, nur aus zuverlässigen Vorhersagen zu lernen.

Unsere Evaluierung auf reellen Daten zeigt, dass selektives Lernen eine wesentliche Verbesserung der Qualität der Empfehlungen im Vergleich zu Systemen ohne selektives Lernen mit sich bringt.

## 2 Selektives Lernen für Empfehlungsmaschinen

In diesem Kapitel definieren wir das selektive Lernen für Empfehlungsmaschinen und schlagen drei Ansätze vor, die diese Definition auf unterschiedliche Arten implementieren.

> **Definition:** *Selektives Lernen* für Empfehlungsmaschinen umfasst Methoden zum Lernen und Vorhersagen von Nutzerpräferenzen nicht aufgrund aller verfügbaren Daten, sondern aufgrund einer Selektion von Daten und Aspekten der Modelle, die die Qualität der Empfehlungen maximieren.

Eine formale Definition befindet sich in der Dissertation, die die Grundlage für diesen Beitrag ist [Ma17a]. Sie kann hier aus Platzgründen nicht ausführlich beschrieben werden.

## 2.1 Selektives Vergessen

Selektives Vergessen erfordert zwei grundlegende Neuerungen bei der strom-basierten Verarbeitung von Ratings (eine Form von Nutzerfeedback). Diese Neuerungen spiegeln sich in den Komponenten einer Empfehlungsmaschine wieder. Die Abbildung 1 bietet eine Übersicht über die Komponenten und die Architektur einer solchen Empfehlungsmaschine.

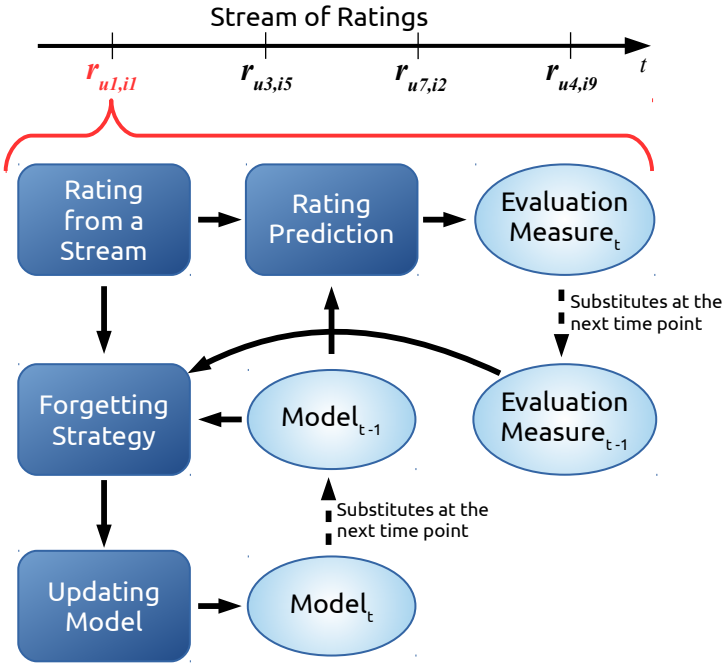



Abb. 1: Übersicht über die Komponenten einer strom-basierten Empfehlungsmaschine mit selektivem Vergessen [Ma17a]

Die erste der neuartigen Komponenten ist eine Vergessensstrategie. Diese Komponente entscheidet, welche Informationen vergessen werden sollen. Dabei muss beachtet werden, dass nicht nur alte Informationen vergessen werden können. Im Kapitel 2.1.1 geben wir ein Beispiel für eine solche Strategie.

Die zweite neuartige Komponente setzt das Vergessen der Information um, die durch eine Vergessensstrategie selektiert wurde. Ein einfaches Löschen der Daten aus einer Datenbank ist nicht ausreichend, da ein Modell bereits von den zu löschenden Daten gelernt hat. Diese Komponente entfernt den Einfluss der zu löschenden Daten aus dem Modell. Sie ist in dem Update-Mechanismus der strom-basierten Algorithmen angesiedelt (s. Abbildung 1). Wir haben drei alternative Algorithmen zur Umsetzung des Vergessens vorgeschlagen und ausführlich evaluiert. Dabei haben wir einen Repräsentanten der Matrix-Faktorisierung-Algorithmen, den BRISMF Algorithmus [Ta09], mit unseren Vergessensmethoden erweitert.

### 2.1.1 Vergessensstrategien

Im Rahmen der Dissertation wurden elf Vergessensstrategien vorgeschlagen, die einer der zwei Kategorien zugeordnet werden können:

1. Rating-basierte Vergesensstrategien
2. Vergessensstrategien im latenten Raum

Aus Platzgründen geben wir hier jedoch nur ein Beispiel für eine solche Vergessensstrategie.

**Sensitivitätsbasiertes Vergessen** Diese Strategie ist ein Beispiel für eine rating-basierte Vergesensstrategie. Sie arbeitet direkt mit Ratings (Nutzerfeedback) und entscheidet für jedes Rating, ob es vergessen oder behalten werden sollte.

Diese Vergessensstrategie basiert auf der Idee der lokalen Sensitivität der Präferenzmodelle. Sobald ein neues Rating beobachtet wurde, verwenden strom-basierte Algorithmen dieses Rating zur Aktualisierung des zugehörigen Nutzerprofils. Infolge dieser Aktualisierung wird das Profil angepasst. Der Grad der Veränderung des Profils kann gemessen werden.

Wenn das gegebene Rating konsistent mit dem bisherigen Nutzerprofil war, dann soll es nur eine geringfügige Anpassung des Modells hervorrufen. Wenn eine große Änderung des Profils beobachtet wurde, dann kann angenommen werden, dass das Rating nicht in die bisherigen Präferenzen des Nutzers passt. Das betroffene Rating könnte ein Ausreißer sein (z.B. ein Geschenk für eine andere Person).

Um solche Ausreißer zu erkennen, kann der Grad der lokalen Veränderung des Nutzermodells gemessen werden. In Matrix-Faktorisierung-Algorithmen, die die State-of-the-art in Empfehlungsmaschinen sind, wird ein Nutzermodell in Form eines Vektors mit latenten Faktoren gespeichert (s. [Ta09] für die Erklärung der Grundlagen zum latenten Raum). Sei $p_u^t$ ein Vektor des Nutzers $u$ zum Zeitpunkt $t$. Mit der folgenden Variable $\Delta_{p_u}$ kann der Unterschied im Modell eines Nutzers zwischen den Zeitpunkten $t$ und $t+1$ im $k$-dimensionalen latenten Raum erfasst werden.

$$\Delta_{p_u} = \sum_{i=0}^{k} (p_{u,i}^{t+1} - p_{u,i}^{t})^2 \tag{1}$$

$\Delta_{p_u}$ kann über die Zeit, im Verlauf eines Datenstroms beobachtet werden. Somit können der Mittelwert $\overline{\Delta_{p_u}}$ und Standardabweichung $SD(\Delta_{p_u})$ ermittelt werden. Eine Veränderung des Nutzermodells gilt dann als abnormal hoch (z.B. bei einem Ausreißer), wenn die folgende Ungleichung gilt:

$$\Delta_{p_u} > \overline{\Delta_{p_u}} + \alpha \cdot SD(\Delta_{p_u}) \tag{2}$$

$\alpha$ ist ein Hyperparameter, der die Sensitivität der Vergessensstrategie kontrolliert. Er wird experimentell so bestimmt, dass die Empfehlungsqualität maximal ist.

### 2.1.2 Ergebnisse zum Selektiven Vergessen

Um den Einfluss von Vergessentmethoden auf die Empfehlungsqualität zu untersuchen, haben wir mehr als 1040 Experimente durchgeführt. Dabei haben wir eine State-of-the-art-Methode mit selektivem Lernen erweitert und haben dann die Erweiterung mit der Standard-Variante verglichen. Die Experimente umfassten eine Offline-Evaluierung mit Testdaten, Hyperparameteroptimierung und statistischen Tests mit dem Friedman-Test und Wilcoxon-Rangsummentest. Die Ergebnisse wurden gegen multiples Testen mit Hommel's Methode korrigiert, um die Alphafehler-Kumulierung zu vermeiden. Als Evaluierungsmaß haben wir das inkrementelle Recall von Cremonesi et al. verwendet [CKT10].

Die Abbildung 2 zeigt eine Auswahl der Ergebnisse unserer Evaluierung (für alle Ergebnisse s. [Ma17a]). Links in den Plots ist die Vergleichsmethode (Matrixfaktorisierung ohne selektives Lernen). Die sonstigen Boxplots zeigen die Ergebnisse der Matrixfaktorisierung mit unterschiedlichen Vergessensstrategien. In unseren Experimenten haben wir eine *signifikante Verbesserung* der Empfehlungsqualität auf sieben von acht Datensätzen beobachtet.

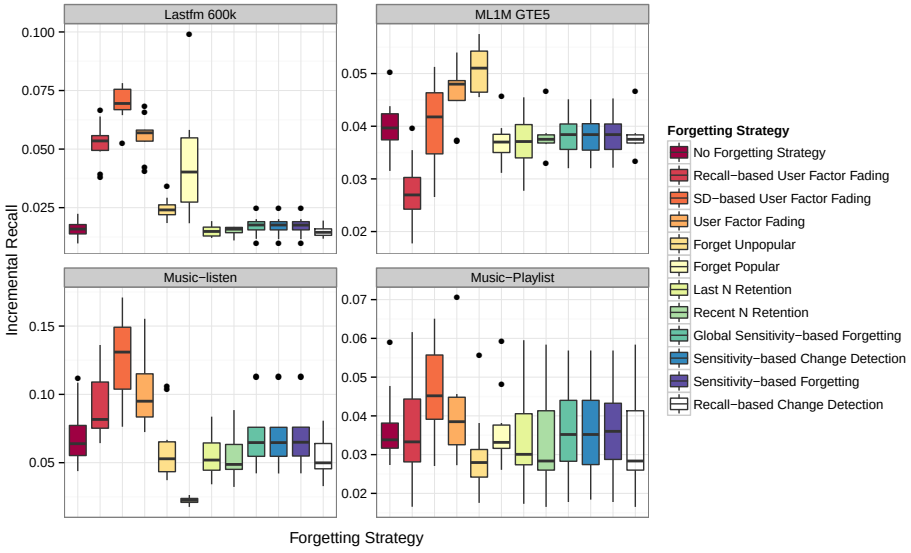



Abb. 2: Inkrementelles Recall einer Empfehlungsmaschine mit und ohne Vergessensmethoden auf vier Datensätzen mit implizitem Nutzerfeedback. Auf jedem Datensatz gibt es Vergessensstrategien, die die Vergleichsbaseline (No Forgetting) übertreffen (höhere Werte sind besser) [Ma17b].

## 2.2 Hoefding-CF

Unser zweite Ansatz zum selektiven Lernen wurde für nachbarschaftsbasierte Collaborative Filtering Methoden (CF) entwickelt. Die CF-Methoden funktionieren auf eine ähnliche Art und Weise wie die konventionellen kNN-Methoden (k Nearest Neighbours) im maschinellen Lernen. Vereinfacht ausgedrückt, suchen sie nach ähnlichen Nutzern und empfehlen, was die ähnlichen Nutzer relevant fanden. Diese Methoden arbeiten oft mit einem Schwellenwert für die Ähnlichkeit zwischen potenziellen Nachbarn und dem aktiven Nutzer.

Wir schlagen ein weiteres Kriterium für eine mehr restriktive Selektion der Nachbarn vor. Dieses Kriterium basiert auf Hoeffding-Bound, einer theoretischen Schranke für eine Abweichung einer Beobachtung vom Mittelwert. Unser Kriterium berücksichtigt nicht nur die absolute Ähnlichkeit zwischen zwei Nutzern, sondern auch die Anzahl der Beobachtungen, die zur Berechnung dieser Ähnlichkeit verwendet wurden. Somit, berücksichtigt unser Kriterium die Zuverlässigkeit von Ähnlichkeiten.

Laut dem von uns vorgeschlagenen Kriterium ist der aktive Nutzer $u_a$ nur dann zu einem weiterem Nutzer $u_x$ zuverlässig ähnlich, wenn die folgende Ungleichung gilt:

$$\widehat{sim}(u_a,u_x) - \widehat{sim}(u_a,u_B) > 2\varepsilon \tag{3}$$

$u_B$ ist ein sogenannter Basline-Nutzer. Für Baseline-Nutzer haben wir mehrere Implementierungsmöglichkeiten vorgeschlagen (s. [Ma17a]). Intuitiv kann man sie als Durchschnittsnutzer oder zufällige Nutzer verstehen. Das bedeutet, dass die Ähnlichkeit zwischen $u_a$ und $u_x$ nur dann zuverlässig ist, wenn sie signifikant höher ist als die Ähnlichkeit von $u_a$ zum Baseline-Nutzer. Der Begriff der Signifikanz und der Wert von $\varepsilon$ wurden von der Hoeffding-Bound hergeleitet. Demnach gilt [Ho63]:

$$\varepsilon = \sqrt{\frac{R^2 \cdot \ln(1/\delta)}{2n}} \tag{4}$$

wo $n$ die Anzahl der Beobachtungen ist, die zur Berechnung der Ähnlichkeit verwendet wurden. $R$ steht für den Wertebereich der Ratings und $\delta$ bestimmt das Signifikanzniveau.

### 2.2.1 Ergebnisse zu Hoefding-CF

Unsere Experimente auf vier reellen Datensätzen zeigen, dass unser selektives Kriterium verwandte Methoden (wie Shrinkage [BKV07] oder Significance Weighting [He99]) übertrifft. Eine grafische Darstellung der Ergebnisse ist in der Abbildung 3 zu sehen.

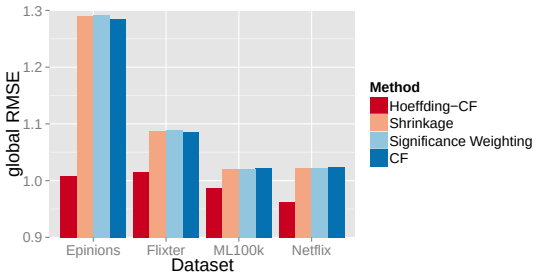



Abb. 3: Ergebnisse der CF-Methoden mit unterschiedlichen Methoden der Nachbar-Selektion [MS14]

## 2.3 Teil-überwachtes Lernen

Unser letzte Ansatz zum selektiven Lernen ist teil-überwachtes Lernen für strom-basierte Empfehlungsmaschinen. In diesem Ansatz lernen Empfehlungsmaschinen von den zahlreichen fehlenden Werten in der Rating-Matrix (nicht bewertete Artikel). Da die Anzahl der fehlenden Werte in der Matrix extrem hoch ist (typischerweise 99%), müssen die Methoden, die diese Information nutzen, selektiv vorgehen. Wir schlagen das erste Framework für teil-überwachtes Lernen für strom-basierte Empfehlungsmaschinen vor.

In unserem Framework vergleichen wir zwei Ansätze zum teil-überwachten Lernen: Co-Training und Self-Learning. Im Co-Training werden mehrere Modelle parallel trainiert. Die Vorhersage eines Modells kann dann als Trainingsbeispiel durch die restlichen Modelle verwendet werden. Auf diese Weise trainieren sich die Modelle gegenseitig, indem sie Vorhersagen füreinander als Trainingsbeispiele bereitstellen. Im Self-Learning ist das Prinzip ähnlich. Anstatt jedoch aus Vorhersagen anderer Modelle zu lernen, lernt das Modell in diesem Fall aus eigenen Vorhersagen.

Das Lernen aus Vorhersagen bringt allerdings Risiken mit sich. Wenn die Vorhersagen inkorrekt sind, dann lernen die Modelle aus falschen Daten. Um dem Problem vorzubeugen, dürfen die Modelle nicht aus allen Vorhersagen lernen, sondern sie handeln *selektiv*. Wir schlagen zahlreiche Methoden zur Auswahl der zuverlässigen Vorhersagen vor, aus denen die Modelle anschließend lernen.

Um das selektive teil-überwachte Lernen auf Datenströmen umzusetzen, haben wir auch weitere neuartigen Komponenten in unserem Framework eingeführt. Die Abbildung 4 stellt eine Übersicht dieser Komponenten dar. Eine der wichtigsten Komponenten ist das Zuverlässigkeitsmaß. Dieses Maß wird genutzt, um die Zuverlässigkeit einer Vorhersage zu schätzen. Diese ist von essenzieller Bedeutung bei der Entscheidung, ob ein Modell aus einer bestimmten Vorhersage lernen soll, oder nicht.

Eine weitere wichtige Komponente ist das Training-Set-Splitter. Diese Komponente sorgt dafür, dass die Modelle, die parallel trainiert werden, unterschiedliche Daten bekommen. Wäre das nicht der Fall, dann wären die Modelle identisch und sie könnten sich gegenseitig

nichts beibringen. Diese und weitere neuartige Komponenten des Frameworks werden im Details in der Dissertationsschrift beschrieben [Ma17a].

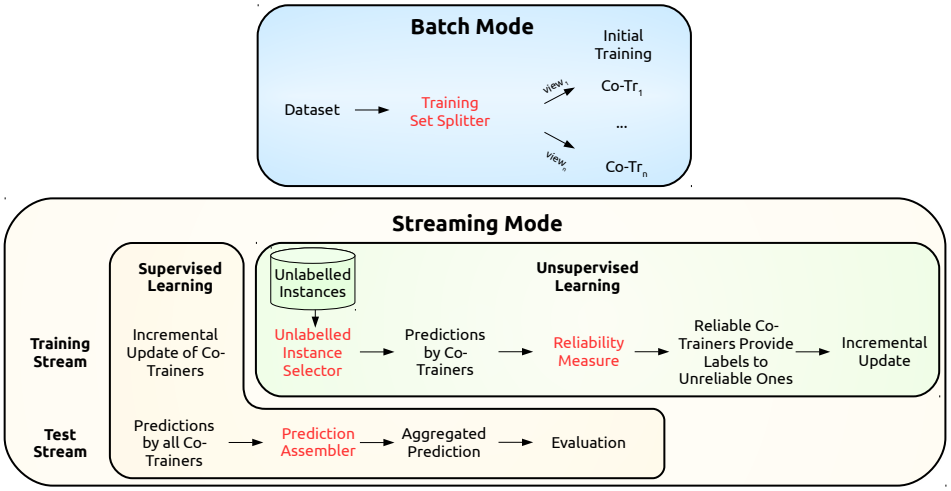



Abb. 4: Übersicht der Komponenten des Frameworks für teil-überwachtes Lernen für strom-basierte Empfehlungsmaschinen [MS17]

### 2.3.1 Ergebnisse zum Teil-überwachten Lernen

Um das selektive Lernen aus Vorhersagen inklusive der zahlreichen Komponenten des neuartigen Frameworks zu evaluieren, haben wir mehr als 700 Experimente durchgeführt. In diesen Experimenten haben wir Co-Training mit zwei (SSL2) und mit drei (SSL3) parallelen Modellen mit Self-Learning (SL) und mit einer Baseline ohne teil-überwachtes Lernen (NoSSL) verglichen.

Die Abbildung 5 stellt die Ergebnisse auf einem der getesteten Datensätze dar. Das Evaluierungsmaß ist das früher erwähnte inkrementelle Recall (höhere Werte sind besser). Aus der Abbildung ist ersichtlich, dass Co-Training mit drei Modellen (SSL3) alle anderen Methoden dominiert. Experimente auf anderen Datensätzen haben ähnliche Ergebnisse geliefert. Unsere statistische Analyse hat ergeben, dass SSL3 auf allen getesteten Datensätzen die Empfehlungsqualität gegenüber NoSSL signifikant verbessert hat. Dies geht jedoch mit einem erhöhten Rechenaufwand einher.

## 3 Zusammenfassung

In dieser Arbeit haben wir eine neues Lern-Paradigma für Empfehlungsmaschinen, *selektives Lernen*, vorgeschlagen. Entgegen der geltenden Annahme, dass alle verfügbaren Daten zum Lernen genutzt werden sollen, haben wir gezeigt, dass eine intelligente Auswahl der Lerndaten zu signifikant besseren Empfehlungen führt.

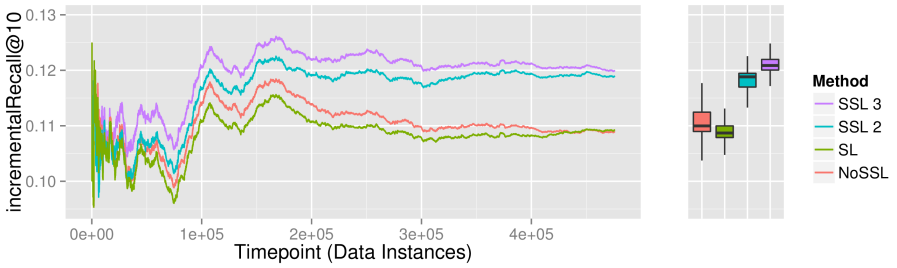



Abb. 5: Inkrementelles Recall der teil-überwachten Methoden (SSL und SL) und der Basline ohne tail-überwachtes Lernen (NoSLL) auf dem Movielens 1M Datensatz (höhere Werte sind besser) [MS17].

Wie haben drei Arten des selektiven Lernens entwickelt: selektives Vergessen von Rating oder Modellteilen, Selektion der Nachbarn im nachbarschaftsbasierten Collaborative Filtering, und Selektion der Vorhersagen für teil-überwachtes Lernen von Nutzerpräferenzen. Bei allen diesen Methoden des selektiven Lernens hat unsere ausführliche Evaluierung mit reellen Daten gezeigt, dass das neue Paradigma die Qualität der Empfehlungen *signifikant verbessert*.

In der Zukunft kann des Weiteren erforscht werden, wie sich die Kombination von mehreren selektiven Methoden in einem System auf die Performance der Empfehlungsmaschinen auswirkt. Da vor Allem das teil-überwachte Lernen zu längeren Rechenzeiten führt, kann in der Zukunft das Potenzial der verteilten Berechnung einzelner Modelle auf unterschiedlichen Rechnern untersucht werden.

## Literaturverzeichnis

[BKV07] Bell, Robert; Koren, Yehuda; Volinsky, Chris: Modeling relationships at multiple scales to improve accuracy of large recommender systems. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. 2007.

[Ch12] Chen, Rung-Ching; Huang, Yun-Hou; Bau, Cho-Tsan; Chen, Shyi-Ming: A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. Expert Systems with Applications, 39(4):3995–4006, 2012.

[CKT10] Cremonesi, Paolo; Koren, Yehuda; Turrin, Roberto: Performance of Recommender Algorithms on Top-n Recommendation Tasks. In: Proceedings of ACM RecSys. RecSys '10. ACM, S. 39–46, 2010.

[He99] Herlocker, Jonathan L.; Konstan, Joseph A.; Borchers, Al; Riedl, John: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, NY, USA, S. 230–237, 1999.

[Ho63] Hoeffding, Wassily: Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc., 58:13–30, 1963.

[Li14] Lin, Zhijie: An empirical investigation of user and system recommendations in e-commerce. Decision Support Systems, 68:111–124, 2014.

[Ma11] Manouselis, Nikos; Drachsler, Hendrik; Vuorikari, Riina; Hummel, Hans; Koper, Rob: Recommender Systems in Technology Enhanced Learning. In (Ricci, Francesco; Rokach, Lior; Shapira, Bracha; Kantor, Paul B., Hrsg.): Recommender Systems Handbook. Springer US, Boston, MA, S. 387–415, 2011.

[Ma17a] Matuszyk, Pawel: Selective Learning for Recommender Systems. Dissertation, 2017.

[Ma17b] Matuszyk, Pawel; Vinagre, João; Spiliopoulou, Myra; Jorge, Alípio Mário; Gama, João: Forgetting techniques for stream-based matrix factorization in recommender systems. Knowledge and Information Systems, Aug 2017.

[MS14] Matuszyk, Pawel; Spiliopoulou, Myra: Hoeffding-CF: Neighbourhood-Based Recommendations on Reliably Similar Users. In (Dimitrova, Vania; Kuflik, Tsvi; Chin, David; Ricci, Francesco; Dolog, Peter; Houben, Geert-Jan, Hrsg.): User Modeling, Adaptation, and Personalization, Jgg. 8538 in Lecture Notes in Computer Science, S. 146–157. Springer International Publishing, 2014.

[MS17] Matuszyk, Pawel; Spiliopoulou, Myra: Stream-based semi-supervised learning for recommender systems. Machine Learning, S. 1–28, 2017.

[PNH15] Paraschakis, D.; Nilsson, B. J.; Holländer, J.: Comparative Evaluation of Top-N Recommenders in e-Commerce: An Industrial Perspective. In: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). S. 1024–1031, Dec 2015.

[SKR99] Schafer, J.; Konstan, J.; Reidl, J.: Recommender Systems in E-Commerce. In: Proceedings of ACM Conference on Electronic Commerce. Denver, Colorado, USA, November 1999.

[Ta09] Takács, Gábor; Pilászy, István; Németh, Bottyán; Tikk, Domonkos: Scalable Collaborative Filtering Approaches for Large Recommender Systems. J. Mach. Learn. Res., 10, 2009.

[WLZ15] Wu, Dianshuang; Lu, Jie; Zhang, Guangquan: A fuzzy tree matching-based personalized e-learning recommender system. IEEE Transactions on Fuzzy Systems, 23(6):2412–2426, 2015.

[WP14] Wiesner Martin; Pfeifer Daniel: Health Recommender Systems: Concepts, Requirements, Technical Basics and Challenges. International Journal of Environmental Research and Public Health, 11(3):2580–2607, feb 2014.

**Pawel Matuszyk** ist ein wissenschaftlicher Mitarbeiter am Lehrstuhl für Knowledge Management and Discovery an der Otto-von-Guericke-Universität in Magdeburg. Er promovierte in 2017 auf dem Gebiet von Machine Learning und Data Science. Zu seinen Forschungsschwerpunkten gehören adaptive Empfehlungsmaschinen, Stream-Mining, teilüberwachtes Lernen und inkrementelle Matrix-Faktorisierung. Seine Master- und Bachelor-Abschlüsse erhielt er von der Otto-von-Guericke-Universität in Wirtschaftsinformatik.

# Event-basiertes maschinelles Sehen für agile Roboter[1]

Elias Müggler[2]

**Abstract:**

Kameras sind sehr nützliche Sensoren für mobile Roboter, weil sie klein, passiv und kostengünstig sind sowie reichhaltige Informationen der Umgebung liefern. Obwohl Kameras erfolgreich in einer Vielzahl von Robotern, wie zum Beispiel in autonomen Fahrzeugen oder Drohnen, verwendet werden, stellen Energiebedarf, Latenz, Dynamikbereich und Bildfrequenz beträchtliche Herausforderungen dar. Bildsequenzen von Kameras enthalten viel Redundanz (zeitlich und räumlich) und sowohl das Aufnehmen wie das Verarbeiten dieser Datenmenge benötigt viel Rechenleistung. Dies limitiert die Betriebszeit mobiler Roboter und definiert einen fundamentalen Zielkonflikt zwischen Energiebedarf und Latenz. Spezialkameras für Hochgeschwindigkeits- und Hochkontrastanwendungen sind teuer, schwer und brauchen zusätzliche Lichtquellen, was deren Anwendung in agilen mobilen Robotern verunmöglicht.

In dieser Dissertation werden *Event-Kameras* als bioinspirierte Alternative untersucht um die Limitationen von Standardkameras zu überwinden. Diese neuromorphischen visuellen Sensoren funktionieren auf komplett andere Weise. Anstatt einer Bildsequenz mit einer konstanten Frequenz zu liefern, senden Event-Kameras nur Informationen von den Pixeln, bei denen sich die Helligkeit signifikant verändert hat. Solche pixelweise Veränderungen nennen wir *Events*, welche mit einem Zeitstempel mit der Genauigkeit von *Mikro*-Sekunden versehen und unmittelbar danach asynchron übermittelt werden. Da nur nicht-redundante Informationen übertragen werden, sind Event-Kameras energieeffizient und in der Lage, sehr schnelle Bewegungen zu erfassen. Damit nehmen sie den Zielkonflikt zwischen Energiebedarf und Latenz direkt in Angriff. Zudem verfügen Event-Kameras über einen Dynamikbereich von über 140 dB (Standardkameras verfügen typischerweise um die 60 dB), weil jedes Pixel selbständig ist. Da das Datensignal einer Event-Kamera fundamental anders ist als dasjenige einer Standardkamera (für welche über die letzten fünfzig Jahren Algorithmen für maschinelles Sehen entwickelt wurden) werden neue Algorithmen benötigt, die mit der asynchronen Funktionsweise klarkommen und die hohe zeitliche Auflösung ausnutzen können.

Diese Dissertation präsentiert Algorithmen für Event-Kameras im Bereich Robotik. Da Event-Kameras neuartige Sensoren sind und kommerziell erst seit 2008 erhältlich sind, ist die Literatur über solche Algorithmen spärlich. Dies erschwert die Handhabung dieser Sensoren, eröffnet aber unzählige Möglichkeiten, die es zu erforschen gilt. Diese Dissertation untersucht die Möglichkeiten von Event-Kameras für fundamentale Probleme der Robotik und des maschinellen Sehens wie zum Beispiel Lokalisierung und Steuerung. Unter anderem bietet diese Dissertation Beiträge zur Roboterlokalisierung mittels Event-Kameras. Dafür wird die Lage (Position und Orientierung) bezüglich einer gegebener Karte der Umgebung geschätzt.

[1] Originaltitel: "Event-based Vision for High-Speed Robotics"

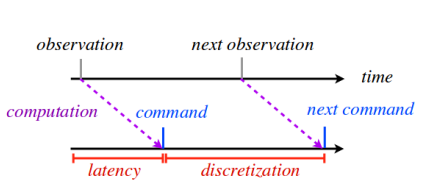
[2] Robotics and Perception Group, Institut für Informatik, Universität Zürich, mueggler@ifi.uzh.ch

# 1 Einführung

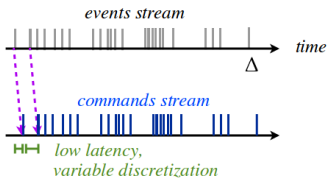
Diese Dissertation beschreibt Algorithmen zur Verarbeitung von Daten von Event-Kameras im Kontext des maschinellen Sehens für mobile und agile Roboter. Im Gegensatz zu gewöhnlichen Kameras, die Bilder mit einer konstanten Rate erfassen und übermitteln, haben Event-Kameras unabhängige, asynchrone Pixel, die lokale Kontraständerungen (sogenannte "Events") zum Zeitpunkt ihres Auftretens übermitteln. Event-Kameras imitieren damit das Funktionsprinzip der Retina [Ma92]. Zusätzlich zu der hohen zeitlichen Auflösung und der geringen Latenz (beide in der Grössenordnung von *Mikro*-Sekunden) bieten Event-Kameras auch einen grossen Dynamikbereich (140 dB im Vergleich zu 60 dB von gewöhnlichen Kameras). Da sich die Ausgabe von Event-Kameras jedoch grundlegend unterscheidet (ein asynchroner Datenstrom von Events anstatt Bildern), sind fundamental neue Algorithmen erforderlich, um mit diesen Daten umzugehen.

Mobile Roboter müssen ihre Umgebung wahrnehmen, um sicher darin navigieren zu können. Damit ein Roboter unabhängig von seiner Umgebung ist, muss er alle Sensoren selbst mitführen. Für agile und schnelle Roboter sind grosse und schwere Sensoren wie z.B. Laserscanner nicht geeignet. Die Verwendung von Kameras als primäre Sensormodalität bietet mehrere Vorteile. Kameras sind klein, leicht und passiv, und benötigen daher nur wenig Strom. Dennoch liefern sie reichhaltige Informationen über die Umwelt. Die Interpretation dieser Daten ist jedoch eine anspruchsvolle Aufgabe, die eine beträchtliche Menge an Rechenressourcen und Rechenleistung in Anspruch nimmt. Zum Vergleich: Mehr als 60 Prozent des menschlichen Gehirns sind an Sehaufgaben beteiligt. In den letzten Jahrzehnten wurden enorme Fortschritte im maschinellen Sehen erzielt, die sogar die menschliche Leistung bei bestimmten Aufgaben übertreffen. Diese Algorithmen werden für Gesichtserkennung und -identifizierung, Bildbeschriftung und viele andere Aufgaben verwendet. Im Bereich der Robotik sind wir jedoch vorerst an der Kartierung der Umgebung und Lokalisierung darin interessiert.

Während gewöhnliche Kameras viele Vorteile bieten, bleiben grosse Herausforderungen bestehen. Erstens leiden Bilder bei hohen Geschwindigkeiten unter Bewegungsunschärfe, die dazu führt, dass nachfolgende Algorithmen versagen. Zweitens bieten Kameras einen eher geringen Dynamikumfang, so dass in Szenen mit grossen Helligkeitsunterschieden bestimmte Bereiche des Bildes unter- oder überbelichtet werden und wichtige Merkmale der Umgebung nicht erkannt werden können. Drittens ist die Bildrate von Kameras begrenzt und bietet somit eine Grenze für die erreichbare Latenzzeit eines Robotersystems (siehe Abb. 1). Viertens: Je höher die Bildrate, desto mehr Daten müssen verarbeitet werden und desto mehr Energie wird für typischerweise unnötige Operationen verbraucht, da aufeinanderfolgende Bilder oft redundant sind. Um diese Einschränkungen zu überwinden, untersuchen wir den Einsatz von Event-Kameras im Bereich der Robotik. Als nächstes wird die Funktionsweise von dieser Kameras beschrieben, deren Vorteile ausgeführt und Herausforderungen diskutiert.




(a) Traditionelle Architektur

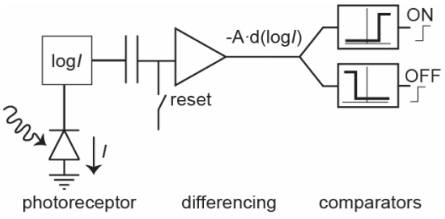



(b) Event-basierte Architektur

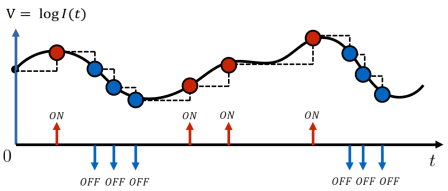
Abb. 1: Vergleich der Ausgabe einer gewöhnlichen Kamera mit einer Event-Kamera, wenn diese für Robotersteuerung verwendet wird: Die geringe Latenzzeit von Event-Kameras ermöglicht eine wesentlich schnellere Reaktion auf Veränderungen in der Umgebung. Abbildungen: [Ce13].

## 1.1 Event-Kameras

Event-Kameras verfügen über unabhängige, asynchrone Pixel, die lokale Helligkeitsänderungen zum Zeitpunkt ihres Auftretens übermitteln. Da sie von der Netzhaut von Wirbeltieren inspiriert sind, werden Event-Kameras auch als "Silizium-Retinas" bezeichnet. Die vereinfachte Schaltung eines einzelnen Pixels ist in Abb. 2a dargestellt. Die Ausgabe ist ein Datenstrom von Events, wenn sich die (logarithmische) Helligkeit um eine benutzerdefinierte Schwelle ändert (vgl. Abb. 2b). Um das Funktionsprinzip des gesamten Sensors zu veranschaulichen, vergleichen wir die Ausgabe einer Event-Kamera mit der einer Standardkamera in Abb. 3 und in einem Video: `https://youtu.be/LauQ6LWTkxM`. Beide Kameras beobachten eine rotierende Scheibe mit einem schwarzen Punkt in der Peripherie. Die Ausgabe einer Standardkamera ist ein Video, d.h. eine Serie von Bildern mit konstanter Frequenz. Stattdessen meldet eine Ereigniskamera nur die Helligkeitsänderungen: Wenn sich nichts verändert, werden auch keine Daten übertragen.




(a) Analoger Schaltkreis (vereinfacht)




(b) Datenstrom (Events) in Abhängigkeit der Helligkeitsveränderung

Abb. 2: Pixel von Event-Kameras. Angepasste Abbildung von [LPD08].

### 1.1.1 Vorteile

**Niedrige Latenzzeit und hohe zeitliche Auflösung.** Event-Kameras bieten eine sehr geringe Latenzzeit für die Erkennung und Übermittlung von Änderungen in der Umgebung, welche in der Grössenordnung von wenigen *Mikro*-Sekunden liegen: Der DVS [LPD08] und der DAVIS [Br14] haben Latenzzeiten von 15 µs bzw. 3 µs. Zum Vergleich: Die La-

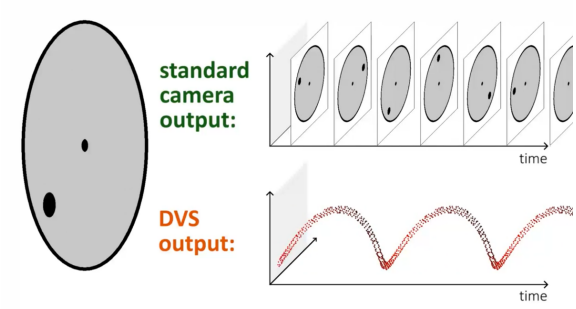



Abb. 3: Vergleich des Datenstroms einer gewöhnlichen Kamera und einer Event-Kamera bei Betrachtung einer rotierenden Scheibe mit einem schwarzen Punkt: Die Standardkamera gibt eine Reihe von Einzelbildern aus, während die Event-Kamera kontinuierlich und asynchron die Änderungen in der Szene meldet. Jeder Punkt in der Raumzeit repräsentiert eine solche Veränderung (sogenannter "Event"). Video-Version: https://youtu.be/LauQ6LWTkxM.

tenzzeit von gewöhnlichen Kameras ist eine gleichmässige Verteilung zwischen 0 und $1/f$, wobei $f$ die zeitliche Diskretisierung (Bildrate) ist. Für eine Standardkamera mit $f = 30\,\text{Hz}$, kann die zeitliche Diskretisierung also bis zu 33 ms sein, was einer um vier Grössenordnungen höheren Latenz entspricht. Daher ermöglichen Event-Kameras wesentlich schnellere Regelkreise (siehe Abb. 1).

**Hoher Dynamikbereich.** Da jedes Pixel einer Event-Kamera unabhängig ist, werden sehr hohe Dynamikbereiche innerhalb der Szene erreicht: Der DVS [LPD08] und der DAVIS [Br14] erreichen Dynamikbereiche von 120 dB resp. 130 dB. Gute Kameras für maschinelles Sehen erreichen typischerweise nur 60 dB. Daher können Event-Kameras auch in Szenen mit sehr hellen und sehr dunklen Bereichen eingesetzt werden (vgl. Abb. 4 für ein Extrembeispiel).

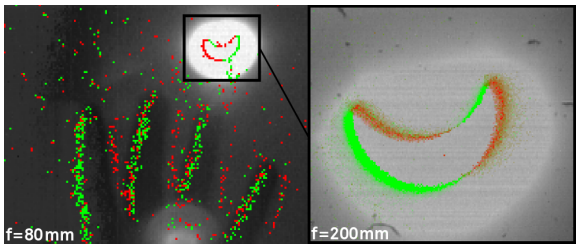



Abb. 4: Eine Sonnenfinsternis, betrachtet mit einer gewöhnlichen und einer Event-Kamera. Während das Bild sowohl über- als auch unterbelichtete Bereiche aufweist, können die Events die Konturen der Finsternis einwandfrei erkennen. Abbildung mit freundlicher Genehmigung von Mark Osswald, Simeon Bamford und Tobi Delbruck.

**Geringe Datenrate.** Da Event-Kameras nur Helligkeits-*Änderungen* auf Pixel-Ebene melden, ist keine Bandbreite erforderlich, wenn ein Pixel den Wert nicht ändert (d.h. wenn es keine Relativbewegung zwischen Szene und Kamera gibt und sich nichts im

Bild bewegt). Darüber hinaus stammen die Events aus kontrastreichen Bereichen (meistens Kanten), so dass viele Vorverarbeitungsoperationen, die für maschinelles Sehen mit gewöhnlichen Bilder notwendig sind, überflüssig werden. Dies ermöglicht eine schnelle Reaktion bei niedriger Rechenleistung (siehe Abb. 1b).

**Tiefer Stromverbrauch.** Eine Event-Kamera benötigt deutlich weniger Strom als eine Standardkamera. Der Hauptgrund dafür ist, dass Analog-Digital-Wandler, die für das Auslesen der Pixel benötigt werden, relativ viel Strom verbrauchen. Solche Konverter sind bei Event-Kameras nicht erforderlich. Deutlich mehr Leistung wird jedoch für die Verarbeitung der Daten verwendet und wenn das gesamte System (Sensorik und Verarbeitung) betrachtet wird, lässt sich deutlich mehr Energie einsparen. Wie Prof. Marc Pollefeys (ETH Zürch) in Bezug auf ein Augmented-Reality-System auf Basis von Standardkameras bemerkte:[3] "Most of the energy is spent moving bits around ... so it would seem natural that ... the first layers of processing should happen in the sensor." (Die meiste Energie wird damit verbraucht, Bits umherzubewegen ... also wäre es logisch, dass ... die ersten Schritte der Verarbeitung im Sensor stattfinden sollten.) Da Event-Kameras bereits den zeitlichen Kontrast berechnen, kann viel Rechenleistung eingespart werden. Wenn die gleiche hohe zeitliche Auflösung (mehrere hundert kHz) benötigt wird, benötigen Standardkameras viel mehr Leistung als Event-Kameras. Solche Hochgeschwindigkeitskameras erfordern in der Regel auch eine starke externe Beleuchtung, die wiederum viel Leistung benötigt.

### 1.1.2 Herausforderungen

Da sich die Datenausgabe von Event-Kameras (ein Datenstrom von Events) grundlegend von der von Standardkameras (eine Folge von Bildern) unterscheidet, können klassische Algorithmen für maschinelles Sehen (Computer Vision) nicht auf Event-Kameras angewendet werden, sodass ein Paradigmenwechsel erforderlich ist. Event-basierte Algorithmen, die die asynchrone Natur der Daten berücksichtigen, müssen entwickelt werden, um die Vorteile von Event-Kameras auszunützen. Aus programmiertechnischer Sicht besteht die Herausforderung darin, sich mit einer anderen Darstellung der visuellen Daten auseinanderzusetzen.

**Asynchrone (Events) statt synchrone (Bilder) Messungen.** Bei Standardkameras basiert die Abtastung auf einem externen Taktgeber, der synchrone Messungen sammelt und ein Bild erzeugt. Im Gegensatz dazu folgen Event-Kameras einer datengesteuerten, adaptiven Abtastrate, d.h. basierend auf Helligkeitsänderungen, die sich in der Szene relativ zur Bewegung des Sensors ergeben, und dies unabhängig für jedes Pixel. Event-Kameras erfüllen daher eine der wichtigsten, impliziten Annahmen der meisten Algorithmen für maschinelles Sehen *nicht*: Die Existenz von Bildern.

[3] http://www.eetimes.com/document.asp?doc_id=1331675

**Helligkeitsunterschiede (d.h. zeitlicher Kontrast) statt absoluter Helligkeit.** Eine weitere Herausforderung besteht darin, dass jeder Event nur binäre Informationen liefert (Helligkeitszu- oder -abnahme, dargestellt durch die Event-Polarität). Das Fehlen von absoluter Helligkeitsinformation scheint jedoch kein Problem zu sein, da nachgewiesen werden konnte, dass sowohl Intensitätsgradienten als auch absolute Intensitäten aus den Events [Co11] rekonstruiert werden können. In den meisten Anwendungen ist keine Rekonstruktion des Bildes erforderlich, da die binäre Darstellung ausreichende Informationen liefert, um eine gegebene Aufgabe zu erfüllen.

**Weitere Herausforderungen.** Aus praktischer Sicht sind einige weitere Herausforderungen zu bewältigen. Das Sensorrauschen heutiger Event-Kameras ist immer noch relativ hoch und ihre Auflösung niedrig ($128 \times 128$ Pixel für den DVS, $240 \times 180$ für den DAVIS). Da die Events asynchron übermittelt werden und die Datenrate von der Kamerabewegung, der Szene, ihrer Textur und den Verzerrungen (Kameraparameter) abhängt, ist die Gewährleistung von Echtzeit-Verarbeitung eine Herausforderung. Die meisten Algorithmen haben eine konstante Rechenzeit *pro Event* und können daher nur bis zu einer bestimmten Event-Rate in "Echtzeit" Ergebnisse berechnen.

## 1.2 Historische Entwicklung von Event-Kameras

Die ersten Silizium-Retinas wurden in den frühen 90er Jahren von Misha Mahowald und Carver Mead [MM89, Ma92, Ma94] entwickelt. Ähnlich wie bei der menschlichen Netzhaut reduziert ihre Silizium-Retina die Datenrate, indem sie die durchschnittlichen Intensitätsniveaus vom Bild subtrahiert und nur räumliche und zeitliche Veränderungen meldet. Mehr als ein Jahrzehnt wurden nur sehr geringe Fortschritte verzeichnet. Wie von [De10] bemerkt, litt die Leistung der frühen Systeme, weil sie gleichzeitig ein neues Paradigma mit kniffliger asynchroner Logik und massiv paralleler analoger Berechnung kombinieren mussten. Im Jahr 2008 wurde schliesslich die erste Event-Kamera kommerziell verfügbar: Der Dynamic Vision Sensor (DVS) [LPD08], der im Rahmen des von der Europäischen Union geförderten CAVIAR-Projekts entwickelt wurde. Ziel des Projektes war der Aufbau eines Event-basierten Hardwaresystems, bestehend aus Sensoren, Verarbeitung, Lernen und Aktuatoren unter Verwendung des Kommunikations-Frameworks *Address-Event Representation* (AER). Das System ermöglicht eine schnelle visuelle Objekterkennung und Tracking-Latenzen in der Grössenordnung von Millisekunden. Der DVS ist ein $128 \times 128$-Pixel-Array, das auf zeitliche Helligkeitsänderungen wie oben beschrieben reagiert und die Sensorebene des CAVIAR-Systems darstellt.

Auf Basis des DVS wurden mehrere Event-Kameras mit zusätzlichen Funktionalitäten entwickelt. Der *Dynamic and Active-pixel Vision Sensor* (DAVIS) [Br14] kombiniert einen DVS mit einer Standardkamera und verwendet dabei die gleichen Pixel. Eine detaillierte Übersicht über Silizium-Retinas ist in [Li15, Chap. 3] zu finden.

## 2 Beiträge dieser Dissertation

Dieses Kapitel fasst die wichtigsten Beiträge der Dissertation zusammen und bringt diese in Zusammenhang. Eine vollständige Liste sowie die kompletten Beiträge können in der Dissertation gefunden werden.[4]

### 2.1 Infrastruktur und Werkzeuge für Event-Kameras

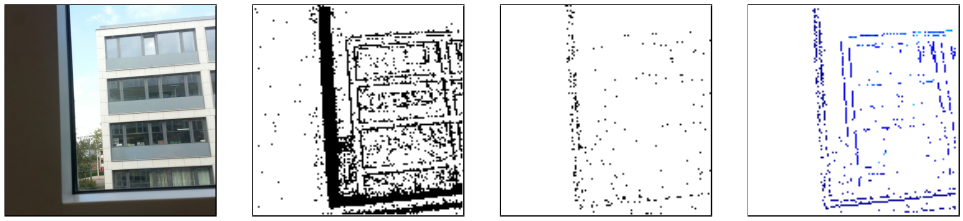
Da Event-Kameras relativ neu sind, gibt es nur wenige öffentlich zugängliche Software. Hauptsächlich wurde ein Java-basiertes Framework (jAER) eingesetzt, das jedoch für schwache Prozessoren, wie sie für Drohnen verwendet werden, zu rechenintensiv ist. Daher wurde zunächst eine ROS-Schnittstelle[5] für die Event-Kameras DVS und DAVIS entwickelt, das zudem ein Tool für intrinsische Kamerakalibrierung und die Charakterisierung der Latenzzeit beinhaltet [MHS14, Mu15a]. Diese Schnittstelle und Tools wurden als Open Source verfügbar gemacht.[6]

Darüber hinaus wurden der erste Event-Kamera-Datensatz[7] und -Simulator[8] veröffentlicht, der sich für Forschung über visuelle und inertiale Schätzung der Kamera-Bahnkurve eignet (wie bewegt sich die Event-Kamera im Raum?) [Mu17]. Dieser Datensatz ermöglicht es Forschern ohne Zugang zu teurer Hardware, Algorithmen zu entwickeln und sie an realen Daten zu testen, und dient als Vergleichsgrundlage zu bereits existierenden Alternativen. Der Datensatz beinhaltet eine grosse Auswahl an verschiedenen Szenen und Bewegungstypen, die in einem Video zu sehen sind: `https://youtu.be/bVVBTQ7l36I`.

### 2.2 Gültigkeitsdauer von Events

Events werden mit einem sehr genauen Zeitstempel versehen, wenn sie generiert werden. Sie enthalten jedoch keine Information über die Dauer, für welche sie eine gültige Messung der Szene sind. Das heisst, dass sie keine Informationen darüber beinhalten, für wie lange der Gradient, der den Event verursacht hat, an der entsprechenden Pixelposition bleibt. Um dieses Problem zu überwinden, führen wir das Konzept von der Gültigkeitsdauer von Events ein [Mu15b], welche als die Zeitdauer beschrieben werden kann, die ein Gradient braucht, um die Distanz von einem Pixel zurückzulegen. Ein Algorithmus, der diese Gültigkeitsdauer aufgrund der Geschwindigkeit des Events berechnet, wurde vorgestellt und evaluiert. Eine direkte Anwendung ist die Erzeugung von scharfen Bildern, die die Kanten der Szene zeigen (siehe Abb. 5).

[4] `http://rpg.ifi.uzh.ch/docs/PhD17_Mueggler.pdf`
[5] Das Robot Operating System (ROS) ist der de-fakto Standard für Robotikanwendungen.
[6] `https://github.com/uzh-rpg/rpg_dvs_ros`
[7] `http://rpg.ifi.uzh.ch/davis_data.html`
[8] `https://github.com/uzh-rpg/rpg_davis_simulator`



(a) Foto der Szene (b) 30 ms (c) 1 ms (d) Gültigkeitsdauer

Abb. 5: Eine Event-Kamera wird diagonal vor einem Fensterrahmen von unten links nach oben rechts bewegt (a). Da der Fensterrahmen viel näher als die Gebäude ist, bewegt er sich im Bild wesentlich schneller. Wenn wir nun ein festes Event-Akkumulationsintervall verwenden, werden die Bilder bei einem zu langen Intervall unscharf (b) oder einige Strukturen sind kaum noch sichtbar, wenn das Intervall zu kurz ist (c). Die präsentierte Methode schätzt die Gültigkeitsdauer jedes Events unabhängig und zeigt den Event nur in diesem Zeitraum an (d).

## 2.3 Position und Orientierung eines Quadrotors

Im Rahmen dieser Dissertation wurde das erste Verfahren entwickelt, das die Schätzung aller sechs Freiheitsgrade eines Quadrotors mit einer Event-Kamera erlaubt [MHS14]. Die Methode ist in der Lage, die Position und Orientierung in Bezug auf ein bekanntes Muster zu bestimmen und dies auch bei schnellen Manövern wie bei einem Salto, bei welchen Rotationsgeschwindigkeiten von mehr als 1200 °/s erreicht werden (siehe Video: `https://youtu.be/LauQ6LWTkxM`). Der Algorithmus verarbeitet die Events asynchron und aktualisiert die Lage des Quadrotors mit sehr geringer Latenzzeit und hoher Genauigkeit.


(a) Quadrotor-Salto. (b) Algorithmus.

Abb. 6: Lage-Berechnung eines Quadrotors während eines Saltos.

## 2.4 Lokalisierung der Kamera und Schätzung der Bahnkurve

Wir stellen einen Event-basierten Ansatz für die Schätzung der Event-Kamera-Bewegungen vor, welcher die Kameralage (Position und Orientierung) beim Eintreffen eines jeden Events berechnet und so die Latenzzeit praktisch eliminiert [Ga17]. Unsere Methode ist

die erste Arbeit, die sich mit Event-basierter Lage-Berechnung mit sechs Freiheitsgraden und Bewegungen in realistischen und natürlichen Szenen beschäftigt. Die Methode ist ebenfalls in der Lage, sehr schnelle Bewegungen zu verfolgen, bei welchen die Bilder von Standardkameras stark verschwommen wären (siehe Video: https://youtu.be/iZZ77F-hwzs). Die Methode wurde sowohl im Innen- als auch im Außenbereich erfolgreich evaluiert.

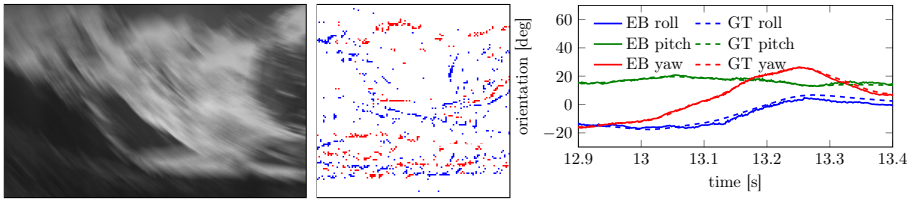



Abb. 7: Sequenzen mit sehr hoher Kamera-Geschwindigkeit. Oben links: Bild einer Standardkamera mit deutlich erkennbarer Bewegungsunschärfe. Oben rechts: Events während eines Intervalls von 3 ms (Farbe zeigt Event-Polarität). Unten: Geschätzte Orientierung der entwickelten Methode (EB) mit kurzer Latenzzeit und hoher zeitlicher Auflösung. Vergleichsdaten eines Motion-Capture-Systems (GT) dienen als Vergleich.

## 3 Ausblick auf zukünftige Forschung

Während sich Algorithmen für Event-basiertes maschinelles Sehen noch in einem frühen Stadium befinden, haben mehrere Publikationen im Rahmen dieser Dissertation ihre Vorteile gegenüber hochwertigen Standardkameras unter Beweis gestellt. Die Leistungsfähigkeit in Bezug auf Präzision und Robustheit muss jedoch weiter erhöht werden, um mit bestehenden Ansätzen, die auf Standardkameras basieren, konkurrieren zu können. Dies erfordert unter anderem verbesserte Event-Kameras (bezüglich Auflösung, Messrauschen, etc.) und weitere Fortschritte aufseiten der Algorithmen.

## Literaturverzeichnis

[Br14] Brandli, Christian; Berner, Raphael; Yang, Minhao; Liu, Shih-Chii; Delbruck, Tobi: A 240x180 130dB 3us Latency Global Shutter Spatiotemporal Vision Sensor. IEEE J. Solid-State Circuits, 49(10):2333–2341, 2014.

[Ce13] Censi, Andrea; Strubel, Jonas; Brandli, Chistian; Delbruck, Tobi; Scaramuzza, Davide: Low-latency localization by Active LED Markers tracking using a Dynamic Vision Sensor. In: IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS). 2013.

[Co11] Cook, Matthew; Gugelmann, Luca; Jug, Florian; Krautz, Christoph; Steger, Angelika: Interacting maps for fast visual interpretation. In: Int. Joint Conf. Neural Netw. (IJCNN). S. 770–776, 2011.

[De10] Delbruck, Tobi; Linares-Barranco, Bernabe; Culurciello, Eugenio; Posch, Christoph: Activity-driven, event-based vision sensors. In: IEEE Int. Symp. Circuits Syst. (ISCAS). S. 2426–2429, Mai 2010.

[Ga17] Gallego, Guillermo; Lund, Jon E. A.; Mueggler, Elias; Rebecq, Henri; Delbruck, Tobi; Scaramuzza, Davide: Event-based, 6-DOF Camera Tracking from Photometric Depth Maps. IEEE Trans. Pattern Anal. Machine Intell., 2017.

[Li15] Liu, Shih-Chii; Delbruck, Tobi; Indiveri, Giacomo; Whatley, Adrian; Douglas, Rodney: Event-Based Neuromorphic Systems. John Wiley & Sons, 2015.

[LPD08] Lichtsteiner, Patrick; Posch, Christoph; Delbruck, Tobi: A 128×128 120 dB 15 $\mu$s latency asynchronous temporal contrast vision sensor. IEEE J. Solid-State Circuits, 43(2):566–576, 2008.

[Ma92] Mahowald, Misha: VLSI Analogs of Neuronal Visual Processing: A Synthesis of Form and Function. Dissertation, California Institute of Technology, Pasadena, California, Mai 1992.

[Ma94] Mahowald, Misha: The Silicon Retina. In: An Analog VLSI System for Stereoscopic Vision. Springer US, Boston, MA, S. 4–65, 1994.

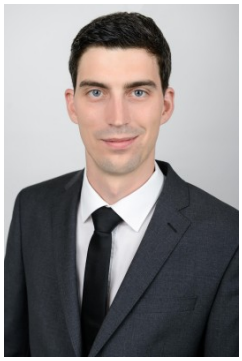
[MHS14] Mueggler, Elias; Huber, Basil; Scaramuzza, Davide: Event-based, 6-DOF Pose Tracking for High-Speed Maneuvers. In: IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS). S. 2761–2768, 2014.

[MM89] Mead, Carver A.; Mahowald, M.A.: A silicon model of early visual processing. Neural Netw., 1(1):91–97, 1989.

[Mu15a] Mueggler, Elias; Baumli, Nathan; Fontana, Flavio; Scaramuzza, Davide: Towards Evasive Maneuvers with Quadrotors using Dynamic Vision Sensors. In: Eur. Conf. Mobile Robots (ECMR). S. 1–8, 2015.

[Mu15b] Mueggler, Elias; Forster, Christian; Baumli, Nathan; Gallego, Guillermo; Scaramuzza, Davide: Lifetime Estimation of Events from Dynamic Vision Sensors. In: IEEE Int. Conf. Robot. Autom. (ICRA). S. 4874–4881, 2015.

[Mu17] Mueggler, Elias; Rebecq, Henri; Gallego, Guillermo; Delbruck, Tobi; Scaramuzza, Davide: The Event-Camera Dataset and Simulator: Event-based Data for Pose Estimation, Visual Odometry, and SLAM. Int. J. Robot. Research, 36:142–149, 2017.



**Elias Müggler** ist Research Scientist bei Oculus Research. Er erhielt seine Bachelor- und Master-Abschlüsse in Maschinenbau von der ETH Zürich in den Jahren 2010 und 2012. Seine Doktorarbeit an der Universität Zürich bei Prof. Dr. Davide Scaramuzza zum Thema Event-Kameras für Robotik schloss er 2017 *summa cum laude* ab. Während seines Studiums hat er Aufenthalte an der Chalmers University of Technology in Göteborg, am Massachusetts Institute of Technology sowie am IIT in Genua absolviert. Seine Forschungsinteressen liegen im Bereich Robotik und Maschinelles Sehen. Er hat in diesen Bereichen 6 Journal- und 16 Konferenzpublikationen veröffentlicht und hat regelmässig für angesehene Journale (TPAMI, JFR, TNNLS) und Konferenzen (ICRA, IROS) Reviews verfasst. Für seine Arbeiten hat er den KUKA Innovation Award 2014, ein Qualcomm Innovation Fellowship 2016 und den Misha Mahowald Prize for Neuromorphic Engineering 2017 erhalten.

# Neue Verfahren gegen das Aufoktroyieren von Verhaltensveränderungen in Software[1]

Stefan Nürnberger[2]

**Abstract:** *Malicious Software*, kurz *'Malware'*, stellt in der heutigen, digitalisierten Welt ein immer größer werdendes Problem dar. Prinzipiell wird jedes IT-System dadurch bedroht, dass sich über Sicherheitslücken Systeme manipulieren lassen, ihren Dienst nicht mehr wie einst programmiert ausführen und dadurch neben finanziellem Schaden auch Gefahr für Leib und Leben darstellen können. Das Ausmaß an Malware ist in den letzten zehn Jahren förmlich explodiert: Nach aktuellen Untersuchungen [AV18] kursieren derzeit über 700.000 bekannte Malwares weltweit – jedes Jahr kommen ca. 100.000 neue hinzu. Aufgrund der enorm hohen Komplexität moderner Software können sich unterschiedlichste Sicherheitslücken einschleichen. Während eine automatische Identifizierung von Sicherheitslücken bisher nur unzureichend zu bewerkstelligen ist, so haben die am häufigsten auftretenden Sicherheitslücken dennoch eine gemeinsame Ursache [MI11]: *Memory Corruption Errors* ermöglichen es Angreifern sowohl Eingaben zu steuern, Berechnungen zu verändern oder Ausgaben zu fälschen, beispielsweise mit dem Ziel Online-Banking-Transaktionen zu ändern, Spam-Email-Server unbemerkt zu installieren oder Nutzer zu erpressen, indem die Festplatten dieser Opfer verschlüsselt werden.

In meiner Dissertation habe ich die Hauptverbreitungsursache bekämpft, indem ich vier neue, sich ergänzende Schutzmechanismen konzipierte, welche gegen *Memory Corruption Errors* schützen. Alle Methoden haben gemein, dass sie existierende Programme schützen können, ohne auf deren Quellcode angewiesen zu sein, der oft in unsicheren Sprachen wie *C* oder *C++* geschrieben wurde. Zum einen verändern die Methoden die Speicherverwaltung des Betriebssystem oder des Prozessors, um bestimmte Angriffsmuster präventiv zu verhindern. Zum anderen aber auch Methoden, die Prozessorinstruktionen zur Laufzeit umschreiben, um die Angriffsfläche auf ein statistisch insignifikantes Niveau zu senken. Und zuletzt auch Methoden, welche zwar den Angriff an sich zulassen, jedoch eine Manipulation zuverlässig erkennen und dadurch deren Auswirkungen im Keim ersticken können. Die vorgestellten Methoden wurden nicht nur konzipiert, sondern auch für das Betriebssystem Linux implementiert und evaluiert.

[1] Englischer Titel der Dissertation: "Mitigating the Imposition of Malicious Behaviour on Code"

[2] Universität des Saarlandes / CISPA, nuernberger@cispa.saarland

# 1 Problemstellung

In den Anfangsjahren handelte es sich bei Computerviren um eigenständige Programme, die unerwünschtes Verhalten zeigten oder gar das System unter ihre Kontrolle brachten. Der Oberbegriff *„Malware“*, in dem das Wort *„Software“* steckt, legt auch nahe, dass es sich dabei um separate Software handelt mit ihrer eigenen Funktionalität. Doch mit fortschreitenden Sicherheitstechniken, die das Ausführen von unbekannter Software verhindern, wurden die Methoden der Angreifer immer raffinierter. Die seit einigen Jahren am weitesten verbreitete Methode ein Rechnersystem zu kompromittieren, ist nicht mehr das Einschleusen einer separaten Software, sondern das gezielte Manipulieren bereits laufender Software. Diese Manipulationen sind keineswegs auf Abwandlungen des ursprünglichen Verhaltens limitiert, sondern mittlerweile so ausgefeilt, dass beliebiges Verhalten, nämlich Turing-vollständige Berechnungen, von außen in einer laufenden Software provoziert werden können. In der Konsequenz bedeutet dies, dass die größte Gefahr darin lauert, dass gutartigen Programmen zur Laufzeit ein anderes Verhalten aufoktroyiert wird, anstatt, wie bisher, dass neue Programme ins System eingeschleust werden.

Während bösartige Programme die klare Absicht des Diebstahls oder der Manipulation von Daten haben, hat ein gutartiges Programm in aller Regel einen Nutzen für den Anwender. Wenn nun aber ein Programmierfehler dazu führen kann, plötzlich das Verhalten eines Programms zu verändern, bleibt dies vor traditionellen Virenscannern verborgen, weil sich am Programmcode selbst nichts verändert. Diese Angriffe sind gemeinhin als *„Code Reuse Attacks’”* [Sc15, Sn13, Da13, BN14] bekannt, weil Instruktionen eines Programms gezielte in falscher Reihenfolge wiederverwendet werden („reuse“) und somit zu beliebigem Verhalten führen. Dies wird dadurch bedingt, dass Instruktionen (Programminhalt) und Programmflusskontrolle voneinander unabhängig sind. Dies ist ein Feature, welches uns erst die Konzepte von *Funktionen* und letztendlich *Rekursion* erlaubt: Der Speicherinhalt des Codes bleibt unverändert, doch die Flusssteuerung des Prozessors kann Instruktionen wieder und wieder in beliebiger Reihenfolge ausführen. Code-Reuse-Angriffe zielen darauf ab nur die Flusskontrolle zu unterwandern und dadurch den Code-Bereich des Speichers gar nicht zu modifizieren: Der Angriff kann deshalb unbemerkt stattfinden. Bereits 1972 stießen Forscher auf diese Art von Angriff, als sie zeigten, dass man durch nicht ordnungsgemäß verarbeitete Eingaben eines Programms dessen Flusskontrolle, und damit dessen Verhalten, beliebig ändern kann [An72, S.61]. Dies ist beispielsweise möglich, indem Kontrollstrukturen überschrieben werden (*Buffer Overflow*). Sicherheitslücken, die zu derartigen Speicherfehlern führen belegen in der Klassifizierung der häufigsten Programmierfehler immer einen der ersten drei Plätze [MI11]. Ist eine solche Lücke erst einmal entdeckt, so hat die massenhafte Verbreitung derselben Software zur Folge, dass ein *Exploit*, d.h. das programmatische Ausnutzen dieser Schwachstelle, millionenfach Anwendung finden kann.

## 2 Lösungen

Den Kern meiner Dissertation prägen vier innovative Sicherheitsmechanismen zur Abwehr von Code-Reuse-Angriffen. Einzeln betrachtet bieten diese Verfahren bereits effektiven und effizienten Schutz. Ihr volles Potential entfalten sie jedoch erst im gegenseitigen Zusammenspiel. Die im Rahmen meiner Dissertation entstandenen Publikationen sind in enger Kooperation mit äußerst geschätzten Kollegen erfolgt. Zur klaren Abgrenzung meiner Beiträge weise ich daher in den nachfolgenden Darstellungen der Einzellösungen explizit auf eben diese Inhalte hin. Die genannten Papiere gehören zu den best-zitiertesten meiner Publikationsliste. Sie trugen wesentlich dazu bei den State of the Art um einen weiteren bedeutsamen Schritt voranzutreiben. Im Folgenden werden die vier Lösungen kurz vorgestellt und in einen zeitlicher Kontext gebettet. Abbildung 1 zeigt die typischen Zusammenhänge bei einem Angriff: Während der Entwicklungsphase enthält der Sourcecode bereits eine Sicherheitslücke, welche später ausgenutzt werden wird. In der aus der Zuverlässigkeitslehre entlehnten Kausalkette *Fault→Error→Failure* wird die Existenz einer solchen Lücke als *Fault* bezeichnet. Wird das kompilierte Programm dann später in den Arbeitsspeicher geladen, kann der Angreifer die Lücke ausnutzen: Der *Error* tritt auf. Auf diese Weise können z.B. Rücksprungadressen manipuliert werden. Auswirkungen hat diese Manipulation *zunächst* keine, bis die eingeschleuste Rücksprungadresse auch verwendet wird (*Failure*). Die hier präsentierten Lösungen setzen an unterschiedlichen Punkten dieser Kette an und sind komplementär zueinander.

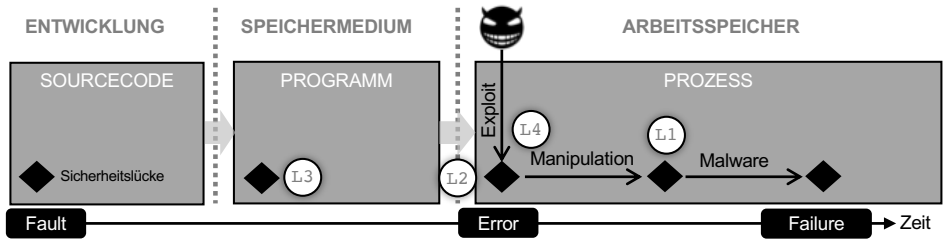



Abb. 1: Die Lösungen L1-L4 im Zusammenhang der Kausalitätskette *Fault→Error→ Failure*

L1**: Kontrollflussintegrität.** Die erste Lösung (Kapitel 2.1) ist so konzipiert, dass sie Exploits nicht abwendet, sondern deren Manipulationen am Kontrollfluss erkennt und verhindert. L1 basiert auf dem Prinzip der Kontrollflussintegrität (*Control Flow Integrity*, kurz *CFI*). Dafür wird aus dem Binärprogramm der Kontrollflussgraph (CFG) extrahiert und zur Laufzeit permanent mit dem gewünschten Kontrollfluss abgeglichen, um eine Manipulation zu erkennen. So kann eine Kompromittierung des Systems vereitelt werden. Der vorgestellte Ansatz war der erste, der Binärprogramme nicht modifiziert und somit kompatibel zu signierten Binärdateien ist, wie sie z.B. in Apples iOS Betriebssystem vorkommen. Er trägt deshalb den Namen *MoCFI* (Mobile CFI). Diese Arbeit wurde 2012 im Rahmen des IEEE-Symposiums *Network and Distributed System Security* (NDSS) [Da12] veröffentlicht.

`L2`**: Instruktionsentropie.** Ein Schutz, der frühzeitiger als `L1` gegen Code-Reuse-Angriffe schützt stellt Kapitel 2.2 vor. `L2` garantiert, dass Exploits gar nicht erst angewandt werden können, indem durch Mutation keine zwei identischen Programme oder Prozesse mehr existieren. Dieser Ansatz macht sich *Dynamic Binary Rewriting* zunutze, eine Technik, die das ad-hoc-Umschreiben von Binärcode erlaubt, um in Echtzeit Sicherheitsmechanismen einzupflanzen und Code zu verändern, sodass ihn kein Angreifer wissen kann. Dazu wird der unveränderte Binärcode von der Festplatte oder SSD geladen, in kleine Stücke unterteilt und diese dann durcheinandergewürfelt, um eine hohe Entropie zu erreichen. Nun genügt es nicht mehr, dass ein Angreifer eine Kopie eines Programms (bestehend aus Code und Daten) besitzt, vielmehr müsste er dessen genaues Speicher-Layout kennen, um es angreifen zu können. Diese Arbeit wurde 2013 auf dem ACM-Symposium *Computer and Communication Security* (AsiaCCS) [Da13] veröffentlicht.

`L3`**: Oxymoron.** `L3` setzt noch früher in der Kausalitätskette an und schreibt unsichere Programme in sichere, randomisierte Programme um (Kapitel 2.3). `L3` löst das Problem, dass der Sicherheitsgewinn durch Mutation dem langjährigen Paradigma von *Shared Memory* widerspricht. *Shared Memory* in Kombination mit *Copy-on-Write* spart eine signifikante Menge an Arbeitsspeicher. Ein randomisiertes Speicherlayout verhindert das Teilen und Wiederverwenden desselben Speichers. Meine Lösung *Oxymoron* stellt die erste Kombination von Methoden dar, die sich eigentlich gegenseitig ausschließenden: *Shared Memory* und *Code-Mutation*. Da bisheriger x86-Code die Kombination von Shared Memory und feingranularer Randomisierung nicht ermöglicht, habe ich eine neue Calling-Convention für x86-Prozessoren konzipiert, die keine absoluten Adressen mehr enthält. Die Arbeit wurde im Rahmen des USENIX Security Symposiums im Jahr 2014 veröffentlicht [BN14].

`L4`**: XnR - Das neue Bit.** Da durch feingranulare Speicherrandomisierung die Speicheradressen vor einem Angreifer nur versteckt werden, sind Angriffe theoretisch möglich, sobald ein Angreifer den isolierten Adressraum eines Prozesses auslesen kann. Einen solchen Angriff stellt das sog. *JIT-ROP (Just-in-Time-Return-oriented Programming)* dar. JIT-ROP Angriffe stützen sich auf eine zweite unwahrscheinliche, aber nicht unmögliche, Sicherheitslücke: Speicherisolations-Fehler. Dadurch können sie eine feingranulare Speicher-Randomisierung auslesen und Exploits konstruieren. Die Lösung `L4` (Kapitel 2.4) verhindert, dass ebensolche Speicherisolations-Schwachstellen ausgenutzt werden können. Zu diesem Zweck habe ich den Virtual-Memory-Mechanismus des Linux Kernels modifiziert, sodass er ein nicht existierendes Hardware-Feature *XnR (eXecute no Read)* emulieren kann. Das von mir als *„XnR“* bezeichnete Speicher-Bit annotiert Speicherbereiche, die ausführbar aber nicht lesbar sein dürfen. Diese bis dato in Prozessoren nicht vorhandene Funktionalität verhindert, dass ein Angreifer Wissen über Code eines randomisierten Prozesses erhält. Diese Arbeit wurde auf der ACM-Konferenz Computer and Communication Security (CCS) im Jahr 2014 [Ba14] veröffentlicht.

### 2.1 L1: Kontrollflussintegrität (CFI)

Die Idee von Kontrollflussintegrität (Control Flow Integrity, CFI) geht auf Abadi et al. [Ab05] zurück. CFI verhindert nicht beabsichtigte Kontrollflüsse, indem zur Laufzeit der aktuell angestrebte Kontrollfluss mit dem beabsichtigtem Kontrollfluss verglichen wird. Mein Beitrag beinhaltete das Entwickeln und Implementieren der Extrahierung des CFGs aus einem entschlüsselten iOS-Programm. Des Weiteren konzipierte und implementierte ich den Laufzeitprüfungsmechanismus, der den tatsächlichen Kontrollfluss mit dem statisch extrahierten Kontrollfluss vergleicht. Zu diesem Zweck wird der Kontrollflussgraph (CFG) aus dem Programm extrahiert. Dabei bilden Folgen von Instruktionen ohne mögliche Kontrollflussveränderung (*Basic Blocks*) die Kanten. Diejenigen Instruktionen hingegen, die eine Kontrollflussveränderung vornehmen können, werden als Knoten dargestellt (Abbildung 2).

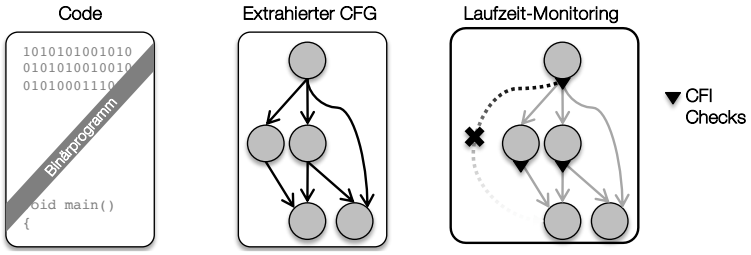



Abb. 2: Der Kontrollflussgraph (CFG) wird aus Binärdateien extrahiert und zur Laufzeit permanent gegen den versuchten Kontrollfluss geprüft.

Für Knoten des CFG werden indirekte Sprünge, indirekte Verzweigungen, Funktionsaufrufe und Funktionsrücksprünge berücksichtigt. Ihre Zieladressen sind zur Laufzeit variabel und müssen daher geschützt werden. Darüber hinaus sind direkte Funktionsaufrufe mit fest codierten Zielen enthalten, damit die Rückkehradresse einer Funktion überprüft werden kann (*Shadow Stack*). Ebenfalls wurde das für Objective-C typische *Message Passing* mit CFI-Methoden abgesichert. Message Passing erlaubt es, Methoden geerbter anderer Objekte aufzurufen. Die Überprüfung des Kontrollflusses zur Laufzeit wird dadurch erreicht, dass ein *Inline-Reference-Monitor (IRM)* alle Instruktionen mit potentiellen Kontrollflussänderungen ersetzt und an zentraler Stelle gegenprüft. Jegliche Abweichung vom zuvor extrahierten CFG führt zu einer Unterbrechung des Programms und verhindert dadurch schadhafte Folgen (siehe Abbildung 3).

Die als *MoCFI* benannte Lösung macht sich auf der Zielplattform iOS sog. *In-Memory-Patches*, d.h. das dynamische Ändern von Code zur Laufzeit, zunutze. Dazu verwendet MoCFI eine Shared Library, die in jeden Prozess geladen wird und beim Starten eines neuen Prozesses die notwendigen Checks an den Kontrollfluss-Instruktionen einbaut. Da solche Checks für gewöhnlich viel mehr Instruktionen beinhalten, als die typischerweise *einzige* Instruktion, die sie ersetzt, wurde mit sog. *Trampolinen* und *Exception Handlers* gearbeitet. Beide sind jeweils eine einzige Instruktion lang und erlauben so das Ersetzen von Code, ohne die Originallänge zu verändern. Dies hat den Vorteil, dass relative und absolute Adressen in Code und Daten erhalten bleiben. Diese Vorgehensweise macht den

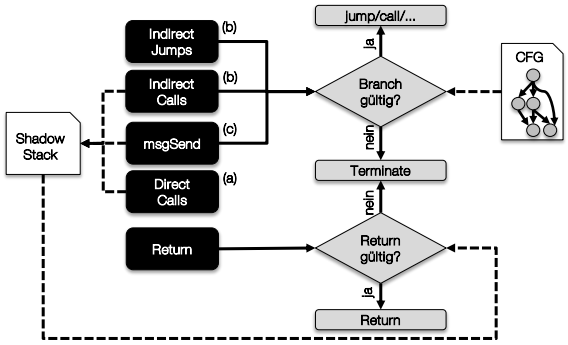



Abb. 3: Der Inline Reference Monitor Ansatz: Sämtliche Möglichkeiten der Kontrollflussänderungs müssen anhand des extrahierten CFGs überprüft werden.

Ansatz mit ASLR, statischen Programmsignaturen und Verschlüsselung kompatibel. Die notwendige Shared Library wird zusammen mit den zu schützenden Prozessen geladen, indem per Jailbreak die Umgebungsvariable `DYLD_INSERT_LIBRARIES` gesetzt wird. Diese gaukelt dem Betriebssystem vor, dass die eingeschleuste Shared Library zum Prozess gehört. Als Prozessorarchitektur kommt die ARM-Architektur zum Einsatz, die der am weitesten verbreite Befehlssatz für mobile Endgeräte (iOS und Android) ist.

Die Lösung `L1` besticht durch die Fähigkeit jegliche Art von Angriff zu erkennen, die das Verhalten eines Programms verändert und ist dadurch agnostisch in Bezug auf die unzähligen Arten von Angriffen. Die durchgeführten Performance-Analysen (Seiten 58ff. der Dissertation) zeigen, dass Performance-Einbußen typischerweise zwischen 1% und 20% rangieren.

### 2.2 `L2`: Instruktions-Entropie durch Binary Rewriting

Eine orthogonale Lösung zu CFI (`L1`) besteht darin, die Ursache von Code-Reuse Angriffen und nicht nur deren Symptome zu schützen. Um erfolgreich zu sein, muss ein Angreifer Adressen im verwundbaren Prozess kennen. Dies kann durch einen *Dangling-Pointer*, eine `printf`-Schwachstelle oder einen Buffer Overflow mit Leserechten geschehen. Eine Methode dem entgegenzuwirken besteht darin, Instruktionen an nicht vorhersagbare Adressen zu verteilen. Die in Betriebssystemen bereits vorhandene Address Space Layout Randomisation (ASLR) bietet hier keinen Schutz, da nur die Basisadresse des Codes verschoben wird. Die als `L2` umgesetzte *feingranulare Randomisierung* hingegen verschiebt jede Instruktion an eine andere Adresse im Speicher. So wird verhindert, dass Code-Adressen bekannt sind, was einen effektiven Schutz gegen Code Reuse Angriffe bietet. Um die Schutzmethode auch ohne Quellcode anwenden zu können, kommt ein ein Binary Rewriter zum Einsatz. Um bei jedem Programmstart randomisieren zu können, muss der Binary Rewriter während des Ladevorgangs des Prozesses alle notwendigen Umschreibungsaufgaben ausführen. Folglich muss dieser Prozess sehr schnell sein, damit der Prozessstart nicht verzögert wird. Den zugrundeliegenden Binary Rewriter habe ich von

Grund auf neu entwickelt und implementiert. Zuvor war kein existierender Rewriter in der Lage, beim Prozessstart bereits ad hoc Instruktionen umzuschreiben. Mein Beitrag umfasste die Konzeption und Implementierung aller notwendigen Algorithmen, wie z.B. die Erstellung der Kontrollflussgraphen, die Darstellung von Instruktionen in einer *Intermediate Language*, die notwendigen Randomisierungsalgorithmen, Optimierungsalgorithmen, das Laden von Prozessen und das Parsen von ausführbaren Dateien und Shared Libraries. Der Rewriter unterteilt den Code in beliebig kleine Blöcke, um deren Position im Speicher frei wählen zu können. Um durch Bewegung von Blöcken den originalen Kontrollfluss nicht zu verändert, wird der implizite Kontrollfluss, d.h. das Hintereinanderausführen zweier Instruktionen, durch Einfügen einer `Jump`-Instruktion zwischen zwei Blöcken erhalten (Abbildung 4).

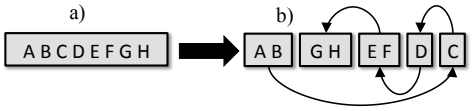



Abb. 4: Der vorhandene Code wird durcheinander gewürfelt und entsprechend der originalen Semantik durch neuen Kontrollfluss miteinander verbunden.

Wann immer es möglich ist, werden Blockgrenzen auf existierende Kontrollfluss-Instruktionen (z. B. *jump*, Funktionsaufruf) im ursprünglichen Programm gelegt. Dieses Vorgehen hat zwei Vorteile: (1) Die explizite Verbindung zweier Blöcke ist bereits vorhanden. (2) Die Laufzeit wird identisch sein, da sich nur das Kontrollflussziel ändert. Meine Analyse hunderter Binärdateien zeigte, dass statistisch gesehen ein aus 1000 Instruktionen bestehendes Programm bereits in 156 Codeblöcke unterteilt. Deren $156! \approx 2^{916}$ Kombinationsmöglichkeiten sind bereits viel höher als die größtmögliche ASLR-Randomisierung auf Byte-Granularität bei einem 64-Bit-System ($2^{64}$). Diese effektive Lösung gegen Code Reuse Attacks erzielte in Benchmark-Ergebnissen (siehe Dissertation S. 92ff) einen Laufzeit-Overhead von 1,2%. Sie profitiert am meisten von einem zusätzlichen Schutz, der in Kaptiel 2.4 vorgestellt wird.

### 2.3 L3: Oxymoron

**Oxymoron** /ˌɒk.sɪˈmɔː.rɒn/ (Substantiv)
*Griechisch: „scharf stumpf". Eine Redewendung, die Widersprüchliches verbindet.*

*Shared Memory* und *Copy-on-Write* stellen sicher, dass mehrfach genutzter Code in Programmen und Libraries nur ein mal Platz im Arbeitsspeicher belegt. Durch die Verwendung größtenteils identischer Libraries und ge`fork`ter Prozesse, werden mehrere GB an Arbeitsspeicher eingespart. Nutzt Code hingegen feingranulare Randomisierung, unterscheiden sich die im Code eingebetteten Adressen: Er ist nicht mehr identisch und müsste in jedem Prozess separat Platz belegen. Zur Lösung dieses Problems habe ich *Position and Layout Agnostic Code (PALACE)* konzipiert. PALACE, eine Untermenge herkömmlichen x86-Codes, vermeidet alle Instruktionen, die absolute oder relative Adressen verwenden. Stattdessen durchlaufen alle Referenzierungen eine indizierte Übersetzungstabelle (siehe Abbildung 5). Der Index beschreibt eindeutig ein Ziel – unabhängig von dessen derzeitiger Adresse. Folglich ändert sich der Speicher für Code nicht, da der semantische Index

unverändert bleibt. Lediglich die Übersetzungstabelle ist in jedem Prozess verschieden. Um *Shared Memory* zu nutzen, teilt *Oxymoron* den Code eines Programms und all seiner Shared Libraries in das kleinste gemeinsam nutzbare Stück: eine Speicherseite (typischerweise 4 KiB). Jede Seite wird dabei in jedem Prozess an einer anderen, zufälligen Adresse eingeblendet. Um herkömmliche Programme in PALACE zu transformieren, habe ich den in Kapitel 2.2 vorgestellten Binary Rewriter erweitert.

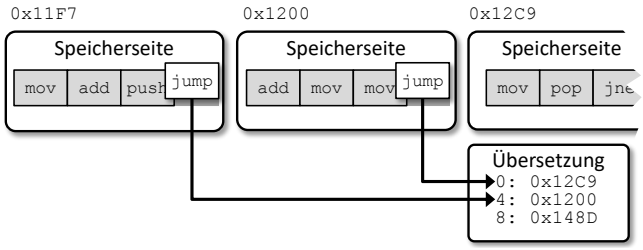



Abb. 5: Speicherseiten werden mit indirekten `jump`-Instruktionen verbunden, deren Zieladresse in der geheimen Übersetzungstabelle steht.

Oxymoron ist von mir so konzipiert, dass (1) die Größe der Übersetzungstabelle minimal ist, (2) die Indirektion effizient ist, (3) die Übersetzung für Angreifer verbogen bleibt und (4) die Lösung auf einem nicht modifizierten Linux-Betriebssystem läuft. Oxymoron verwendet die x86-Funktion der Speichersegmentierung, um die Übersetzungstabelle vor Angreifern zu verbergen. Da es sich bei der Segmentierung um ein Hardwarefeature handelt, ist es naturgemäß sehr effizient. Dies wird von den Benchmarks (Seiten 115-127 der Dissertation) belegt, die einen Overhead von nur 2,7% nachweisen. Somit macht Oxymoron die feingranulare Randomisierung praktikabel und breitflächig anwendbar.

### 2.4 `L4`: XnR

Mithilfe von *Memory Disclosure Attacks* kann ein Angreifer beliebigen Prozessspeicher lesen und so der Kette des Kontrollflusses folgen, um alle Code-Adressen zu erfahren (JIT-ROP). Werden diese Code-Adressen disassembliert können automatisiert nützliche Code-Schnipsel (ROP-Gadgets) für einen Angriff gefunden werden. Ungeachtet der *Memory Disclosure* Schwachstelle, ist dies möglich, weil auszuführender Code im Speicher immer auch lesbar bleiben muss. Das neue „Execute-no-Read“ (XnR)-Bit legt fest, dass ausführbarer Code von Instruktionen nicht mehr gelesen werden kann und bietet dadurch eine Lösung für Memory Disclosure Attacks. Da moderne Prozessoren alle eine Von-Neumann-Architektur aufweisen, die Code und Daten mischt, ist die für XnR notwendige Unterscheidung komplizierter. Die heutige x86-Hardware unterstützt die für XnR benötigten Funktionen nicht, deshalb habe ich diese in Software als Linux-Kernel-Modul implementiert. Das Kernel-Modul blendet absichtlich alle Speicherseiten bis auf die zuletzt benutzte aus, um beim Ausführen von Code per `Page Faults` von Lesezugriffen der CPU zu erfahren.

Wenn die CPU versucht während eines `Instruction Fetch` (siehe Abbildung 6) Speicher zu lesen, stellt dies eine legitime Operation dar und der Code wird weiter fortgesetzt.

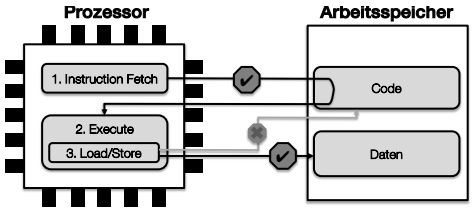



Abb. 6: XnR prüft den Zieltyp von `Load/Store`-Instruktionen.

Ansonsten muss es sich um eine `Load` Instruktion handeln. In diesem Fall muss XnR unterscheiden, ob die Adresse, auf die zugegriffen wird, tatsächlich in einem Bereich liegt, der Daten enthält, oder ob versucht wird aus dem Code zu lesen. In diesem Fall wird der Vorgang abgebrochen und so ein Angriff verhindert. Damit eine ausführbare Datei signalisieren kann, dass sie von XnR geschützt werden möchte, muss das von Linux verwendete Programmformat *ELF* nicht verändert werden. ELF unterstützt die Angabe von Zugriffsberechtigungen für einen bestimmten Speicherbereich bereits. Code, der durch XnR geschützt werden soll, muss lediglich das `R`-Bit (Readable) aus dem ELF-Header des jeweiligen Code-Abschnitts entfernen. Die XnR Kernel-Erweiterung schützt dann vor Memory Disclosure Angriffen und weist dabei nur einen Performance-Overhead von 2,2% auf.

## 3 Fazit

Für die derzeit mehr als 700.000 kursierenden Schadprogramme gilt: sie benötigen eine Methode sich einzunisten, um zur Ausführung zu kommen. Die am weitesten verbreitete Methode ist nach wie vor das Ausnutzen von *Memory Corruption* Sicherheitslücken. Genau dort setzen die sich gegenseitig ergänzenden Lösungen meiner Dissertation an: sie verhindern das *Einnisten* von Schadcode und somit dessen weitere Verbreitung. Das Potential ist dementsprechend groß die Gesamtsituation zu verbessern und die Ausbreitung von Malware stark einzudämmen, anstatt nur partielle, auf bestimmte Schadsoftware zugeschnittene, Lösungen maßzuschneidern. Dabei bieten die Lösungen `L2`, `L3` und `L4` eine abwärtskompatible Methode bestehende Programme im Feld zu schützen.Methoden der statischen Analyse von Quellcode sind extrem wichtig auf dem Weg zu einer allgemein sichereren Software, und in Einzelfällen sogar beweisbar sicheren Software. Sie benötigen jedoch nicht nur Quellcode, sondern auch aktive Mithilfe der Entwickler. Ich verfolge mit meinen Lösungen hingegen einen pragmatischen Ansatz fehlerhafte Programme nachträglich abzusichern. Sie sind somit nicht auf Quellcode angewiesen und ermöglichen die sofortige Absicherung beliebiger Software. Die Lösung `L1` fungiert als zusätzliche Phalanx der Sicherheit, indem Manipulationen selbst nach erfolgreichem Ausnutzen einer Lücke erkannt und verhindert werden.

## Literaturverzeichnis

[Ab05] Abadi, Martín; Budiu, Mihai; Erlingsson, Úlfar; Ligatti, Jay: Control-flow integrity. In: ACM Conference on Computer and Communications Security (CCS). 2005.

[An72] Anderson, James P: Computer Security Technology Planning Study. Bericht, U.S. Air Force Electronic Systems Division (AFSC), Bedford, Massachusetts, 1972.

[AV18] AV-TEST GmbH: , Malware Statistics and Trend Report 2018. `https://www.av-test.org/de/statistiken/malware/`, 2018.

[Ba14] Backes, Michael; Holz, Thorsten; Kollenda, Benjamin; Koppe, Philipp; Nürnberger, Stefan; Pewny, Jannik: You Can Run but You Can't Read: Preventing Disclosure Exploits in Executable Code. In: ACM conference on Computer and communications security (CCS). 2014.

[BN14] Backes, Michael; Nürnberger, Stefan: Oxymoron - Making Fine-Grained Memory Randomization Practical by Allowing Code Sharing. In: USENIX Security Symposium. 2014.

[Da12] Davi, Lucas; Dmitrienko, Alexandra; Egele, Manuel; Fischer, Thomas; Holz, Thorsten; Hund, Ralf; Nürnberger, Stefan; Sadeghi, Ahmad-Reza: MoCFI: A Framework to Mitigate Control-Flow Attacks on Smartphones. In: Symposium on Network and Distributed System Security (NDSS). 2012.

[Da13] Davi, Lucas; Dmitrienko, Alexandra; Nürnberger, Stefan; Sadeghi, Ahmad-Reza: Gadge Me if You Can: Secure and Efficient Ad-Hoc Instruction-Level Randomization for x86 and ARM. In: ACM SIGSAC Symposium on Information, Computer and Communications Security (ASIACCS). 2013.

[MI11] MITRE: , Common Weakness Enumeration Top 25. `http://cwe.mitre.org/top25/`, 2011.

[N7] Nürnberger, Stefan: Mitigating the Imposition of Malicious Behaviour on Code. Dissertation, Universität des Saarlandes, 2017.

[Sc15] Schuster, Felix; Tendyck, Thomas; Liebchen, Christopher; Davi, Lucas; Sadeghi, Ahmad-Reza; Holz, Thorsten: Counterfeit Object-oriented Programming - On the Difficulty of Preventing Code Reuse Attacks in C++ Applications. In: IEEE Symposium on Security and Privacy (S&P). 2015.

[Sn13] Snow, Kevin; Monrose, Fabian; Davi, Lucas; Dmitrienko, Alexandra; Liebchen, Christopher; Sadeghi, Ahmad-Reza: Just-in-time code reuse: On the effectiveness of fine-grained address space layout randomization. In: IEEE Symposium on Security and Privacy. 2013.

**Stefan Nürnberger** wurde am 7. April 1986 geboren. Nach dem Abitur an der Hohen Landesschule in Hanau, begann er 2005 das lang ersehnte Informatikstudium an der TU Darmstadt. Zur Wissenschaft und dem damit verbundenen selbstbestimmten Arbeiten hingezogen, stand die Entscheidung fest nach dem Studium als wissenschaftlicher Mitarbeiter an der Uni zu forschen. Im Jahr 2013 entschied sich Stefan an der Universität des Saarlandes ein Doktorstudium anzuschließen, das er 2017 mit Auszeichnung abschloss. In der Zwischenzeit sammelt er zusätzlich Erfahrung als Teamleiter am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) in Saarbrücken und ist mittlerweile seit 2017 *Faculty* am neu entstandenen Helmholtz-Zentrum für Cybersicherheit CISPA.

Steffen Hölldobler et. al. (Hrsg.): Ausgezeichnete Informatikdissertationen 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2018 191

# Data Profiling – Effiziente Entdeckung Struktureller Abhängigkeiten[1]

Thorsten Papenbrock[2]

**Abstract:** Daten sind nicht nur in der Informatik, sondern auch in allen anderen wissenschaftlichen Disziplinen ein unverzichtbares Wirtschaftsgut. Sie dienen dem Austausch, der Verknüpfung und der Speicherung von Wissen und sind daher unverzichtbar in Forschung und Wirtschaft. Leider sind Daten häufig nicht ausreichend dokumentiert um sie direkt nutzen zu können – es fehlen Metadaten, welche die Struktur und damit Zugriffsmuster der digitalen Informationen beschreiben. Informatiker und Experten anderer Disziplinen verbringen daher viel Zeit damit, Daten strukturell zu analysieren und aufzubereiten. Da die Suche nach Metadaten jedoch eine hoch komplexe Aufgabe ist, scheitern viele algorithmische Ansätze schon an kleinen Datenmengen.

In dieser Dissertation stellen wir daher drei neuartige Entdeckungsalgorithmen für wichtige und zugleich schwierig zu findende Typen von Metadaten vor: Eindeutige Spaltenkombinationen, funktionale Abhängigkeiten und Inklusionsabhängigkeiten. Die vorgeschlagenen Algorithmen übertreffen deutlich den bisherigen Stand der Technik in Laufzeit und Ressourcenverbrauch und ermöglichen so die Nutzbarmachung von erheblich größeren Datensätzen. Da die Anwendung solcher Algorithmen für fachfremde Nutzer nicht einfach ist, schlagen wir außerdem das Programm METANOME vor, das ein praktisches Werkzeug zur Datenanalyse darstellt. METANOME macht dabei nicht nur die in dieser Arbeit vorgeschlagenen Algorithmen nutzbar, sondern auch Entdeckungsalgorithmen für andere Typen von Metadaten. Am Anwendungsfall der Schema-Normalisierung zeigen wir schließlich, wie die effektive Nutzung der gefundenen Metadaten erfolgen kann.

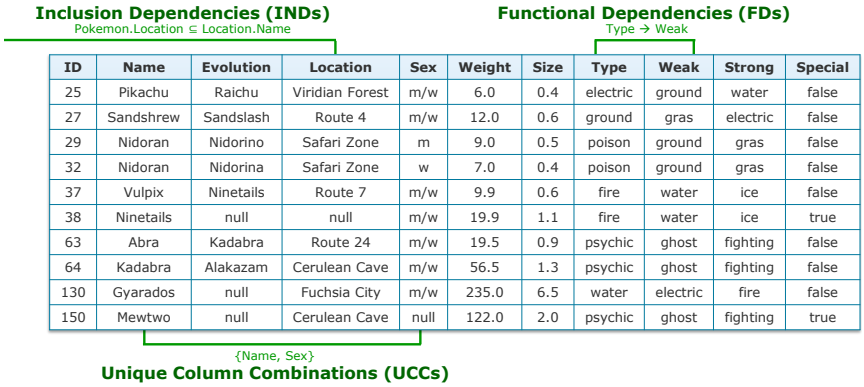
## 1 Extraktion struktureller Metadaten

Data Profiling ist eine Disziplin der Informatik, in der Datensätze mit dem Ziel analysiert werden, deren Metadaten zu bestimmen. Die verschiedenen Typen von Metadaten reichen von einfachen Statistiken wie Tupelzahlen, Spaltenaggregationen und Wertverteilungen bis hin zu weit komplexeren Strukturen, insbesondere Inklusionsabhängigkeiten (INDs), eindeutige Spaltenkombinationen (UCCs) und funktionale Abhängigkeiten (FDs). Sofern vorhanden dienen diese Statistiken und Strukturen dazu, die Daten zu verstehen, sie effizient zu speichern, zu lesen und zu ändern. Da die meisten Datensätze ihre Metadaten aber nicht explizit als beschreibendes Regelwerk zur Verfügung stellen, sind Informatiker häufig gezwungen diese strukturellen Regeln mittels Data Profiling zu bestimmen.

Während einfache Statistiken noch relativ schnell zu berechnen sind, stellen die komplexeren Strukturen schwere, zumeist NP-vollständige Entdeckungsaufgaben dar. In der Regel ist es daher auch mit gutem Domänenwissen nicht möglich, sie händisch zu bestimmen. Es

[1] Englischer Titel der Dissertation: “Data Profiling – Efficient Discovery of Dependencies”

[2] Universität Potsdam, Hasso-Plattner-Institut, Informationssysteme, Prof.-Dr.-Helmert-Str. 2-3, D-14482 Potsdam, Deutschland thorsten.papenbrock@hpi.de




| ID | Name | Evolution | Location | Sex | Weight | Size | Type | Weak | Strong | Special |
|---|---|---|---|---|---|---|---|---|---|---|
| 25 | Pikachu | Raichu | Viridian Forest | m/w | 6.0 | 0.4 | electric | ground | water | false |
| 27 | Sandshrew | Sandslash | Route 4 | m/w | 12.0 | 0.6 | ground | gras | electric | false |
| 29 | Nidoran | Nidorino | Safari Zone | m | 9.0 | 0.5 | poison | ground | gras | false |
| 32 | Nidoran | Nidorina | Safari Zone | w | 7.0 | 0.4 | poison | ground | gras | false |
| 37 | Vulpix | Ninetails | Route 7 | m/w | 9.9 | 0.6 | fire | water | ice | false |
| 38 | Ninetails | null | null | m/w | 19.9 | 1.1 | fire | water | ice | true |
| 63 | Abra | Kadabra | Route 24 | m/w | 19.5 | 0.9 | psychic | ghost | fighting | false |
| 64 | Kadabra | Alakazam | Cerulean Cave | m/w | 56.5 | 1.3 | psychic | ghost | fighting | false |
| 130 | Gyarados | null | Fuchsia City | m/w | 235.0 | 6.5 | water | electric | fire | false |
| 150 | Mewtwo | null | Cerulean Cave | null | 122.0 | 2.0 | psychic | ghost | fighting | true |

Abb. 1: Eine relationale Tabelle mit Daten über Pokémon, die drei ausgewählte Schlüsselbeziehungen zeigt: ein potentieller Primärschlüssel (UCC), ein Fremdschlüssel (IND) und eine innere Schlüsselabhängigkeit (FD).

wurden daher bereits verschiedenste Profiling Algorithmen entwickelt, um die Entdeckung zu automatisieren. Keiner dieser Algorithmen kann allerdings Datensätze von heutzutage typischer Größe verarbeiten, weil entweder der Ressourcenverbrauch oder die Rechenzeit effektive Grenzen überschreitet.

In dieser Arbeit stellen wir neuartige Profiling Algorithmen vor, die automatisch die drei populärsten Typen komplexer Metadaten entdecken, nämlich UCCs, FDs und INDs. Die Popularität dieser drei Strukturen begründet sich in der Tatsache, dass mit ihrer Hilfe die wichtigsten Formen von Schlüssel-Abhängigkeiten beschrieben werden: UCCs beschreiben Schlüssel *für* eine relationale Tabelle, FDs beschreiben Schlüssel *innerhalb* einer relationalen Tabelle und INDs beschreiben Fremdschlüsselbeziehungen *zwischen* relationalen Tabellen. Sie dienen damit nicht nur der Identifikation von Entitäten in einem Datensatz, sondern auch der Verknüpfung, Bereinigung, Anfrage, Integration und logischen Formatierung von Daten. Abbildung 1 zeigt eine Beispielrelation über Pokémon Daten – kleine “Taschenmonster” für Kinder. Von den ebenfalls dargestellten Schlüsselabhängigkeiten lernen wir, dass Pokémon über ihren Namen und ihr Geschlecht eindeutig identifiziert werden (siehe UCC), der Typ eines Pokémon auch dessen Schwäche bestimmt und zusätzliche Informationen über die Herkunft eines Pokémon in einer speziellen anderen Tabelle gefunden werden können.

Die Aufgabe eines Entdeckungsalgorithmus ist es alle gültigen Vorkommen einer Schlüssel-Abhängigkeiten aus einer gegebenen relationalen Instanz zu extrahieren â- ein induktives Such- und Prüfverfahren also, welches die Daten systematisch auf Schlüsseleigenschaften untersucht. Die von uns entwickelten Algorithmen nutzen dazu sowohl bewährte Entdeckungstechniken aus verwandten Arbeiten des Data Profiling, als auch für das Data Profiling neuartige Techniken, wie beispielsweise Teile-und-Herrsche Verfahren, hybrides Suchen, Progressivität, Speichersensibilität, Parallelisierung und innovative Streichungsregeln. Über eine Reihe systematischer Experimente zeigen wir, dass die vorgeschlagenen Algorithmen nicht nur um Größenordnungen schneller sind als alle verwandten Algorith-

men, sie heben auch einige der aktuellen Beschränkungen auf, da sie in der Lage sind Datensätze von häufig vorkommender Größe, d.h. mehrerer Gigabyte Größe, mit akzeptablem Speicher- und Zeitverbrauch zu verarbeiten. Um die entwickelten Algorithmen der Forschungs- und Entwicklungsgemeinschaft, sowie Informatik-Laien zugänglich zu machen, haben wir alle Verfahren in das praktische Profiling Werkzeug METANOME[2] integriert, welches frei und quelloffen verfügbar ist. Zusammengefasst leistet diese Arbeit daher die folgenden Beiträge:

**1. HYFD:** Ein effizienter Algorithmus zur Entdeckung funktionaler Abhängigkeiten, der ein hybrides Suchverfahren zur Identifikation valider FDs im Suchraum einsetzt [PN16].
**2. HYUCC:** Ein effizienter Algorithmus zur Entdeckung eindeutiger Spaltenkombinationen, der ebenfalls auf eine hybride Suche setzt, um Schlüsselkandidaten zu finden [PN17].
**3. BINDER:** Ein effizienter Algorithmus zur Entdeckung von Inklusionsabhängigkeiten, der mittels Datenpartitionierung die Prüfung von IND-Kandidaten beschleunigt [Pa15c].
**4. METANOME:** Ein leicht erweiterbares Data Profiling Werkzeug, das verschiedene Entdeckungsalgorithmen praktisch nutzbar macht [Pa15a].
**5. NORMALIZE:** Ein Algorithmus zur Schema-Normalisierung, der entdeckte Schlüssel-Abhängigkeiten automatisiert bewertet und zur Schema-Reorganisation einsetzt [Pa17].

Im Folgenden erläutern wir zunächst die theoretischen Grundlagen für diese Arbeit und fassen anschließend die einzelnen Beiträge der Dissertation kapitelweise zusammen. Die vollständige Fassung der Dissertation ist in englischer Sprache erschienen [Pa17].

## 2 Schlüsselabhängigkeiten

Das relationale Datenmodell stellt Daten in tabellarischer Form mittels eines festen Schemas dar und ist damit das zur Zeit am weitesten verbreitete Modell. Jede Spalte hat eine eindeutige Bezeichnung, und jede Zeile beschreibt eine in der Tabelle gespeicherte Entität. Spalten und Zeilen werden häufig auch als Attribute und Einträge bezeichnet. Wir definieren nun unsere drei Schlüsselabhängigkeiten als Beziehungen zwischen verschiedenen Spalten:

**Funktionale Abhängigkeiten** (FDs) werden geschrieben als $X \rightarrow A$ und drücken damit aus, dass alle Paare von Einträgen mit gleichem Wert in der Attribut-Menge $X$ auch den gleichen Wert im Attribut $A$ haben – die $X$-Werte bestimmen funktional die $A$-Werte. Wenn wir die relationale Instanz mit $r$ und ihr Schema mit $R$ bezeichnen, dann definiert sich dieser Zusammenhang formal als $\forall t_i, t_j \in r : t_i[X] = t_j[X] \Rightarrow t_i[A] = t_j[A]$, wobei $X \subseteq R$ und $A \in R$ ist. Um alle funktionalen Abhängigkeiten eines Datensatzes zu bestimmen reicht es aus, nur minimale Abhängigkeiten aufzuzählen, bei denen aus der Attributmenge $X$ kein Attribut entnommen werden kann, ohne die FD zu verletzen. Da das Hinzufügen beliebiger Attribute auf der linken Seite der FD die Abhängigkeit nicht verletzt, lassen sich alle nicht-minimalen FDs aus der vollständigen Menge aller minimalen FDs leicht ableiten.

[2] www.metanome.de

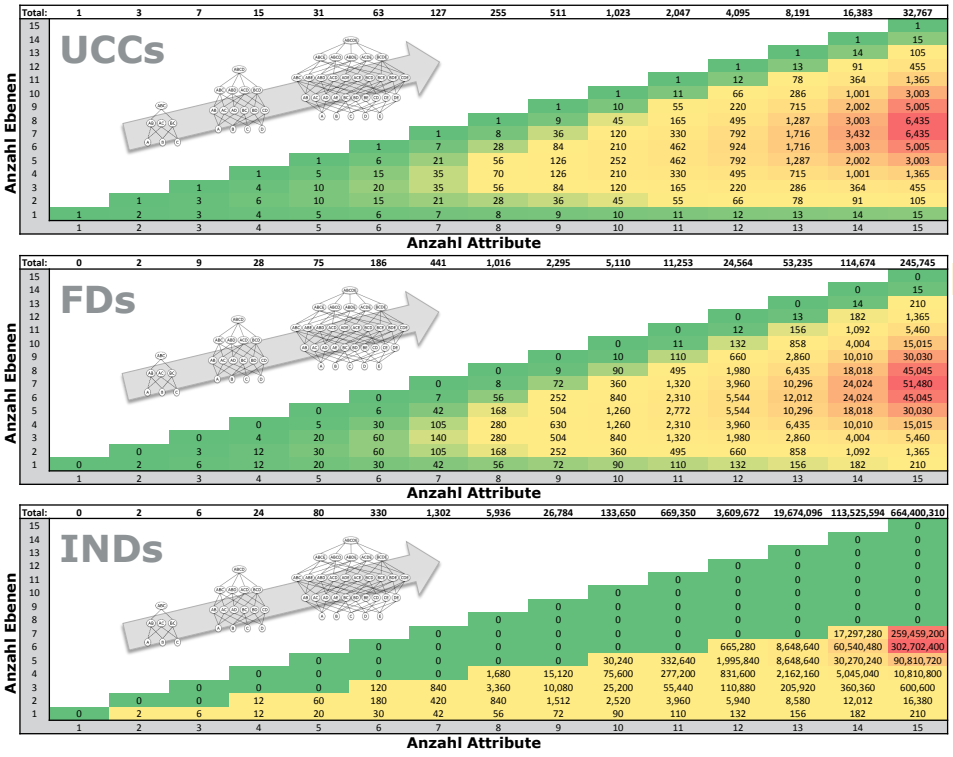



Abb. 2: Wachstum der Suchräume für UCCs, FDs und INDs mit der Anzahl Attribute.

**Eindeutige Spaltenkombinationen** (UCCs) sind Mengen von Attributen $X$, in denen kein Wert (bzw. keine Wertekombination) doppelt vorkommt. Jeder einzelne Wert identifiziert daher eindeutig einen bestimmten Eintrag in der Tabelle. Formal definieren wir eindeutige Spaltenkombinationen $X$ mit $X \subseteq R$ als $\forall t_i, t_j \in r, i \neq j : t_i[X] \neq t_j[X]$. Analog zu FDs reicht es zur Entdeckung aller UCCs aus nur solche aufzuzählen, die minimal sind. Eine minimale eindeutige Spaltenkombination ist jene, die bei Entfernung eines beliebigen Attributes aus $X$ ungültig wird. Aus den minimalen UCCs lassen sich dann ebenfalls alle anderen UCCs ableiten, indem diese durch weitere Attribute ergänzt werden.

**Inklusionsabhängigkeiten** (INDs) werden geschrieben als $R_i[X] \subseteq R_j[Y]$ und besagen, dass alle Werte (bzw. Wertekombinationen) der Attribute $X$ auch in der Menge der Werte von $Y$ vorkommen – sie sind also darin enthalten. Nützlich ist dieser Zusammenhang, weil über Join-Operationen die Einträge, die an diesen Werten hängen, miteinander verbunden werden können. Eine formale Definition dieses Zusammenhangs ist $\forall t_i[X] \in r_i, \exists t_j[Y] \in r_j : t_i[X] = t_j[Y]$, wobei $r_i$ und $r_j$ relationale Instanzen der Schemata $R_i$ und $R_j$ darstellen. Im Gegensatz zu den meisten anderen Arten von Datenabhängigkeiten suchen wir bei INDs nach maximalen Ausdrücken, da beim Entfernen von Attributen auf linker und rechter Seite der Beziehung die Gültigkeit immer erhalten bleibt, die Gültigkeit aber verloren gehen kann, wenn Attribute hinzugefügt werden.

Die Mengen aller FD, UCC und IND Kandidaten wird bestimmt durch die Potenzmengen ihrer linken Seiten. Der Suchraum wird daher häufig als Gitter (engl. lattice) von Kandidaten modelliert: Beginnend mit Beziehungen zwischen einzelnen Attributen werden sukzessiv mehr Attribute zu diesen Kandidaten hinzugefügt. Abbildung 2 visualisiert wie diese Gitter, also die Suchraumgrößen, für die einzelnen Abhängigkeiten mit der Anzahl an Attributen im Datensatz immer weiter wachsen. Bezüglich der Anzahl von Attributen $m$ liegen die Komplexitäten der automatischen Entdeckung dieser drei Arten von Metadaten daher in $\mathcal{O}(2^m)$ für UCCs, $\mathcal{O}(2^m \cdot (\frac{m}{2})^2)$ für FDs und $\mathcal{O}(2^m \cdot m!)$ für INDs (wobei $m!$ eine Vereinfachung ist) [Li12].

## 3 Entdeckung von funktionalen Abhängigkeiten

Um effizient alle minimalen funktionalen Abhängigkeiten zu finden haben verwandte Arbeiten bereits zwei unterschiedliche Ansätze vorgeschlagen: Der erste Ansatz testet die FD Kandidaten im Suchraumgitter systematisch von unten nach oben und streicht dabei als ungültig erkennbare Kandidaten von der weiteren Suche [Hu99]; der zweite Ansatz vergleicht alle Paare von Einträgen in der Tabelle, berechnet aus diesen Vergleichen alle Verletzungen der FDs (entsprechend ihrer Definition) und leitet am Ende aus der Menge aller Verletzungen die Menge aller gültigen minimalen FDs ab [FS99]. In einer evaluierenden Arbeit [Pa15b] haben wir festgestellt, dass der erste Ansatz auf breiten Datensätzen (ca. 40 Attribute und aufwärts) ineffizient wird und der zweite Ansatz bei langen Datensätzen (ca. 100.000 Einträge und aufwärts) Schwächen aufweist, viele Datensätze aber sowohl breit als auch lang sind. Wir schlagen daher den hybriden Algorithmus HyFD vor, der beide Strategien so kombiniert, dass sie sich gegenseitig unterstützen, um sowohl auf breiten als auch langen Datensätzen effizienter zu sein als je eine Strategie für sich allein.

Abbildung 3 veranschaulicht die hybride Suchstrategie von HyFD: In der `Preprocessor` Komponente wird der Datensatz zunächst in kompakte Indexstrukturen übersetzt: Positionslisten-Index (Pli) und komprimierter Entitäten-Index (PliRecord). Anschließend beginnt die `Sampler` Komponente bestimmte Zeilen in der Tabelle zu vergleichen und daraus Verletzungen der FDs zu berechnen. Dies ist eine der beiden Suchstrategien, die allerdings abgebrochen wird, sobald eine überwiegende Mehrheit an Zeilenvergleichen keine neuen Verletzungen mehr geliefert hat – die Strategie ist ineffizient geworden. Daraufhin leitet der `Inductor` des HyFD Algorithmus all jene FD Kandidaten ab, die unter Berücksichtigung der bisher gefundenen Verletzungen noch minimal und gültig sein können. Das Ergebnis formt nun ein Kandidatenset, welches die `Validator` Komponente systematisch mit dem Gitter-Verfahren zu überprüfen beginnt. Auch hier gilt nun, dass HyFD die Validierung dynamisch abbricht, sobald eine Überzahl an Überprüfungen ein negatives Ergebnis liefern und sich die Validierung als ineffizient herausstellt. Wir tauschen nun Vergleichsvorschläge mit der `Sampler` Komponente aus und wechseln zurück zur ersten Suchstrategie. Alle Ergebnisse, die vom `Validator` bisher als gültig identifiziert wurden, sind exakte, minimale funktionale Abhängigkeiten und werden ausgegeben. Der Algorithmus terminiert, sobald dem `Validator` keine weiteren Kandidaten mehr vorliegen.

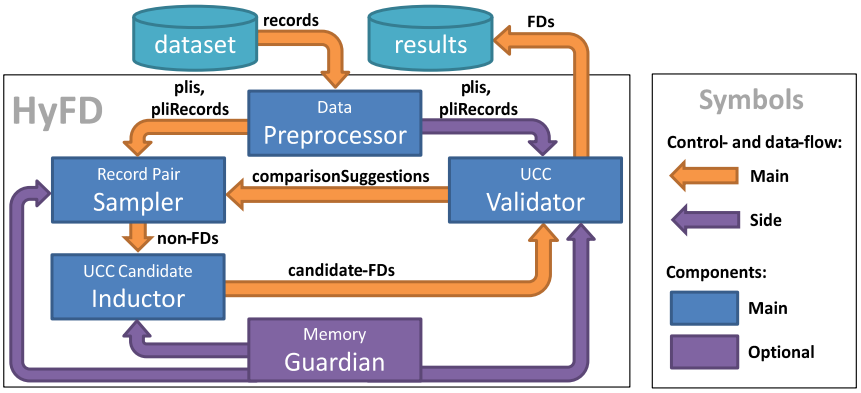



Abb. 3: Übersicht über den HyFD Algorithmus.

Tabelle 1 zeigt die Laufzeiten von HyFD für verschiedene Echtwelt-Datensätze. Die aufgeführten Datensätze bereiten den bisherigen Entdeckungsalgorithmen große Schwierigkeiten, da diese Ansätze entweder ein Speichervolumen von mehr als 100 GB überschreiten oder auch innerhalb mehrerer Tage kein vollständiges Ergebnis berechnen können. Der hybride Ansatz konnte jede Analyse auf einem Rechner mit 16 Prozessorkernen und 128 GB RAM (wovon tatsächlich aber nur ein Bruchteil genutzt wird) problemlos durchführen. Vor allem aber zeigt sich in diesen (und weiteren) Experimenten, dass die Laufzeit mehr von der zu findenden Ergebnisgröße abhängt als von der Größe des Datensatzes – eine ideale Laufzeiteigenschaft für Profiling Algorithmen.

| Datensatz | Spalten [#] | Zeilen [#] | Größe [MB] | FDs [#] | HyFD [s/m/h/d] |
|---|---|---|---|---|---|
| TPC-H.lineitem | 16 | 6 m | 1,051 | 4 k | 4 m |
| PDB.ATOM_SITE | 31 | 27 m | 5,042 | 10 k | 64 m |
| SAP_R3.ILOA | 48 | 45 m | 8,731 | 16 k | 8 h |
| SAP_R3.CE4HI01 | 65 | 2 m | 649 | 2 k | 10 m |
| NCVoter.statewide | 71 | 1 m | 561 | 5 m | 31 h |
| UCI.flight | 109 | 1 k | 1 | 982 k | 54 s |

Tab. 1: Laufzeiten des HyFD Algorithmus auf verschiedenen Datensätzen.

## 4 Entdeckung von eindeutigen Spaltenkombinationen

Zur Entdeckung eindeutiger Spaltenkombinationen nutzt der HyUCC Algorithmus eine sehr ähnliche Strategie wie der HyFD Algorithmus: Nach der Komprimierung des Datensatzes in Indexstrukturen wechseln sich eine Suchstrategie, die auf dem Vergleich von Einträgen basiert, und eine Suchstrategie, die Kandidaten im Suchgitter vergleicht, gegenseitig ab. Die Strategien tauschen Zwischenergebnisse untereinander aus und schlagen zudem ein paar Optimierungen wie beispielsweise das frühzeitige Abbrechen und Parallelisieren von Kandidaten-Checks vor. So ist auch HyUCC allen bisherigen Ansätzen der UCC Entdeckung, die auch jeweils immer nur auf einer Suchstrategie beruhen, überlegen.

Mit HyUCC konnten wir zeigen, dass hybrides Suchen nicht nur für funktionale Abhängigkeiten, sondern auch für viele weitere Arten von komplexen Metadaten eine vielversprechende Strategie ist. So haben sich weitere Entdeckungsalgorithmen mit demselben Suchprinzip angeschlossen, nämlich Aid-FD [Bl16b] für approximative funktionale Abhängigkeiten (approximate FDs), MvdDetector [Dr16] für sogenannte Multivalued Dependencies (MVDs), Hydra [Bl16a] für Denial Constraints (DCs) und HyMD [Dr18] für Matching Dependencies.

## 5 Entdeckung von Inklusionsabhängigkeiten

Die größte Herausforderung bei der Suche nach Inklusionsabhängigkeiten in einem relationalen Datensatz ist die effiziente Ausführung einer extrem großen Anzahl von Kandidaten-überprüfungen. Jede einzelne Überprüfung eines IND Kandidaten ist dabei aufwendiger als die Überprüfung eines FD oder UCC Kandidaten, weil nicht nur die Positionen gleicher Werte in einer Spalte relevant sind, sondern das exakte Übereinstimmen von Werten in verschiedenen Spalten. Eine Komprimierung des Datensatzes in Positionslisten ist daher nicht möglich. Um die Kandidaten möglichst effizient zu prüfen setzen einige IND Entdeckungsalgorithmen – wie auch unser Binder Algorithmus – auf das simultane Testen mehrerer Kandidaten. Das von Binder vorgeschlagene Testverfahren ähnelt einem großen Hash-Join aller Attribute, wohingegen Konkurrenzalgorithmen wie Spider [BLN06] auf Sort-Merge-Joins setzen.

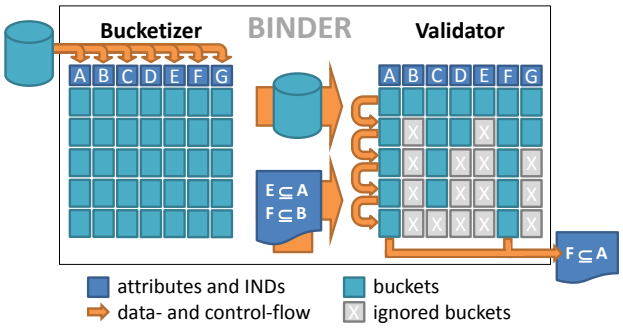



Abb. 4: Bucketierung und Validierung im Binder Algorithmus.

Solch simultane Validierungsverfahren für INDs sind recht effizient, allerdings ist auch der benötigte Speicherbedarf hoch. Damit der Algorithmus nicht – wie einige verwandte Verfahren – an Speichermangel scheitert, schlägt Binder einen Teile-und-Herrsche Ansatz für die Datenhandhabung vor: Wie in Abbildung 4 veranschaulicht wird der Datensatz zunächst Attribut-weise per Hash-Verfahren in Körbe aufgeteilt. Der Algorithmus entfernt dabei doppelte Werte in den Körben, um diese möglichst klein zu halten. Alle Körbe landen schließlich auf der Festplatte und werden dann zum Validieren der IND Kandidaten schrittweise wieder eingelesen. Wurde ein Attribut von allen IND Kandidaten entfernt, so müssen dessen Buckets in kommenden Validierungsschritten nicht mehr mitgelesen werden. Alle Kandidaten, die alle Validierungsschritte überstehen, sind gültige INDs und werden von Binder ausgegeben.

Unsere Experimente zum BINDER Algorithmus zeigen, dass dieser meist effizienter ist als seine Konkurrenz. Der wichtigste Beitrag besteht aber darin, dass der Algorithmus an keiner der bisher bekannten Grenzen scheitert: Er setzt nicht das Vorhandensein einer Datenbank voraus, er scheitert nicht an unzureichend großem Hauptspeicher und er erschöpft nicht das Maximum offener Dateizeiger (engl. file handles) eines Betriebssystems.

## 6 Das Metanome Data Profiling Werkzeug

Aufgrund der praktischen Relevanz des Data Profilings hat die Industrie verschiedene Profiling Werkzeuge entwickelt, die Informatiker in ihrer Suche nach Metadaten unterstützen sollen. Diese Werkzeuge bieten zwar eine gute Unterstützung für die Berechnung einfacher Statistiken, und sie sind auch in der Lage einzelne Abhängigkeiten zu validieren. Allerdings mangelt es ihnen an Funktionen zur echten *Entdeckung* von Metadaten. Um diese Lücke zu schließen schlagen wir das Werkzeug METANOME vor, eine erweiterbare Profiling Plattform, die nicht nur unsere eigenen Algorithmen, sondern auch viele weitere Algorithmen anderer Forscher integriert. Derzeit bieten wir 21 Entdeckungsalgorithmen für 9 verschiedene Arten von Metadaten an und arbeiten kontinuierlich mit Wissenschaftlern anderer Institute, darunter Institute in Frankreich, Italien, Kanada, Neuseeland, Indien, China und den USA, an weiteren Verfahren. Wir haben METANOME ebenfalls in Kooperation mit den Unternehmen FlixBus, IBM und SAP erfolgreich eingesetzt.

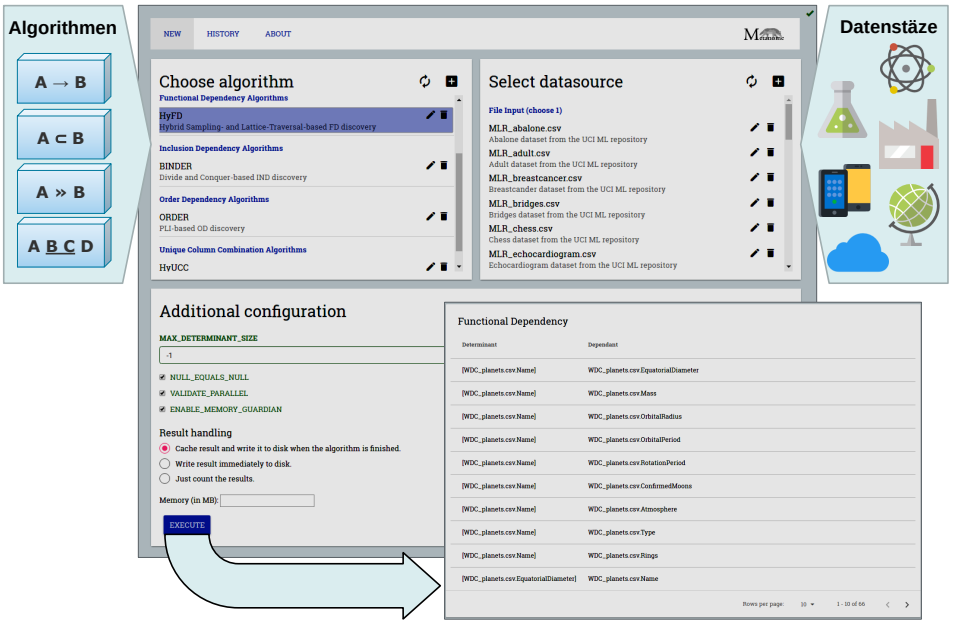



Abb. 5: Die Nutzersicht auf das METANOME Werkzeug mit dem Fenster zur Erstellung von Profiling-Läufen (Mitte) und einem Ergebnis-Fenster (Vorne).

Abbildung 5 zeigt die Nutzeroberfläche des Tools: Im linken Teil der Oberfläche werden die zu entdeckenden Metadaten über ihre Algorithmen ausgewählt, links daneben befindet sich der Import- und Auswahlbereich für Datensätze, und im unteren Teil können Profiling Prozesse konfiguriert und ausgeführt werden. Die Ergebnisse werden dann nach erfolgreicher Ausführung in einer weiteren Maske angezeigt. Auf diese Weise machen wir mit METANOME unsere Forschungsergebnisse für alle Informatiker und auch fachfremde Nutzer zugänglich. Neben der Metadaten-Entdeckung bietet die Plattform auch Unterstützung bei der Bewertung und Visualisierung gefundener Metadaten.

## 7 Metadaten getriebene Schema-Normalisierung

Unsere neuen Profiling Algorithmen ermöglichen die effiziente Entdeckung aller syntaktisch korrekten Metadaten auf realistisch großen Datenstzen. Dies führt nun zur Folgeaufgabe, aus den gefundenen Abhängigkeiten nur die für einen gegebenen Anwendungsfall semantisch bedeutsamen Teile zu extrahieren. Das Extrahieren bedeutsamer Metadaten aus allen findbaren Abhängigkeiten ist eine Herausforderung, da zum einen die Mengen der gefundenen Metadaten überraschenderweise groß sind (oft größer als der untersuchte Datensatz selbst) und zum anderen die Entscheidung über die Anwendungsfall-spezifische semantische Relevanz einzelner Abhängigkeiten schwierig ist.

Um zu zeigen, dass die Vollständigkeit der Metadaten sehr wertvoll für ihre Nutzung ist, veranschaulichen wir die effiziente Verarbeitung und effektive Bewertung von Schlüssel Abhängigkeiten am Anwendungsfall der Schema-Normalisierung: Das NORMALIZE Verfahren bewertet automatisch alle gefundenen Abhängigkeiten daraufhin, ob sie auch semantisch korrekte Schlüssel darstellen und strukturiert die zugehörige Tabelle anschließend entsprechend um. Wir haben NORMALIZE getestet, indem wir zunächst wohlgeformte Schemata denormalisiert haben, dann alle Schlüsselabhängigkeiten mit METANOME finden ließen und anhand dieser Abhängigkeiten schließlich ein normalisiertes Schema abgeleitet haben. Da das abgeleitete Schema dem ursprünglichen sehr ähnlich ist, schließen wir auf eine hohe Effektivität für den NORMALIZE Algorithmus.

## Literaturverzeichnis

[Bl16a] Bleifuß, Tobias: Efficient Denial Constraint Discovery. Masterarbeit, Hasso-Plattner-Institute, Prof.-Dr.-Helmert-Str. 2-3, D-14482 Potsdam, 2016.

[Bl16b] Bleifuß, Tobias; Bülow, Susanne; Frohnhofen, Johannes; Risch, Julian; Wiese, Georg; Kruse, Sebastian; Papenbrock, Thorsten; Naumann, Felix: Approximate Discovery of Functional Dependencies for Large Datasets. In: Proceedings of the International Conference on Information and Knowledge Management (CIKM). S. 1803–1812, 2016.

[BLN06] Bauckmann, Jana; Leser, Ulf; Naumann, Felix: Efficiently Computing Inclusion Dependencies for Schema Discovery. In: ICDE Workshops. S. 2, 2006.

[Dr16] Draeger, Tim: Multivalued Dependency Detection. Masterarbeit, Hasso-Plattner-Institute, Prof.-Dr.-Helmert-Str. 2-3, D-14482 Potsdam, 2016.

[Dr18] Draeger, Tim: Efficient Discovery of Matching Dependencies. Masterarbeit, Hasso-Plattner-Institute, Prof.-Dr.-Helmert-Str. 2-3, D-14482 Potsdam, 2018.

[FS99] Flach, Peter A; Savnik, Iztok: Database dependency discovery: a machine learning approach. AI Communications, 12(3):139–160, 1999.

[Hu99] Huhtala, Ykä; Kärkkäinen, Juha; Porkka, Pasi; Toivonen, Hannu: TANE: An efficient algorithm for discovering functional and approximate dependencies. The Computer Journal, 42(2):100–111, 1999.

[Li12] Liu, Jixue; Li, Jiuyong; Liu, Chengfei; Chen, Yongfeng: Discover Dependencies from Data – A Review. IEEE Transactions on Knowledge and Data Engineering (TKDE), 24(2):251–264, 2012.

[Pa15a] Papenbrock, Thorsten; Bergmann, Tanja; Finke, Moritz; Zwiener, Jakob; Naumann, Felix: Data Profiling with Metanome. Proceedings of the VLDB Endowment, 8(12):1860–1863, 2015.

[Pa15b] Papenbrock, Thorsten; Ehrlich, Jens; Marten, Jannik; Neubert, Tommy; Rudolph, Jan-Peer; Schönberg, Martin; Zwiener, Jakob; Naumann, Felix: Functional Dependency Discovery: An Experimental Evaluation of Seven Algorithms. Proceedings of the VLDB Endowment, 8(10):1082–1093, 2015.

[Pa15c] Papenbrock, Thorsten; Kruse, Sebastian; Quiané-Ruiz, Jorge-Arnulfo; Naumann, Felix: Divide & Conquer-based Inclusion Dependency Discovery. Proceedings of the VLDB Endowment, 8(7):774–785, 2015.

[Pa17] Papenbrock, Thorsten: Data Profiling - Efficient Discovery of Dependencies. doctoralthesis, Universität Potsdam, 2017.

[PN16] Papenbrock, Thorsten; Naumann, Felix: A Hybrid Approach to Functional Dependency Discovery. In: Proceedings of the International Conference on Management of Data (SIGMOD). S. 821–833, 2016.

[PN17] Papenbrock, Thorsten; Naumann, Felix: A Hybrid Approach for Efficient Unique Column Combination Discovery. In: Proceedings of the Conference Datenbanksysteme in Büro, Technik und Wissenschaft (BTW). S. 195–204, 2017.

**Thorsten Papenbrock** wurde 1987 in Mettingen geboren und erlangte dort im Juni 2007 das Abitur am Kardinal-von-Galen-Gymnasium. Anschließend begann er sein Informatik Studium am Hasso-Plattner-Institut an der Universität Potsdam. Er erhielt den Bachelor-Abschluss im Sommer 2010 und den Master Abschluss im Winter 2013. Während seiner Studienzeit absolvierte er Praktika bei der SAP AG in Waldorf (2009; zwei Monate), BBF GmbH in München (2010; zwei Monate) und SAP AG in Belfast (2011; 6 Monate). Ab April 2013 arbeitete er als wissenschaftlicher Mitarbeiter bei Prof. Naumann am Lehrstuhl für Informationssysteme. Neben seiner Forschungstätigkeit, die zu den hier beschriebenen Ergebnissen führte, hat er sich auch in der Lehre in nunmehr 29 Veranstaltungen engagiert. Seit der Einreichung seiner Doktorarbeit im Juni 2017 forscht und lehrt er als Postdoc am Hasso-Plattner-Institut. Neben seinem Forschungsinteresse in den Bereichen Data Profiling, Datenreinigung und verteiltes Rechnen ist er begeisterter Schwimmer und Jogger.

# Momentenbasierte Verfahren für schnelle, transiente Bildgebung und Echtzeitschatten[1]

Christoph Peters[2]

**Abstract:** Wir wenden die Theorie der Momente auf Probleme des Visual Computings an. Aus dieser Theorie entwickeln wir effiziente Algorithmen, die eindimensionale Verteilungen durch eine geschlossene Form aus ihren Momenten rekonstruieren. Solche Rekonstruktionen nutzen aus, dass die ursprünglichen Verteilungen keine negativen Massen beinhalten. Dadurch können sie vor allem bei Verteilungen, die um wenige Punkte lokalisiert sind, mit wenigen Momenten außerordentlich gute Rekonstruktionen erreichen. Wir wenden diese Verfahren auf Messwerte von AMCW Lidar Systemen an. So erhalten wir für jeden Pixel dieser Lichtlaufzeitkameras eine vollständige Rekonstruktion der Impulsantwort des Lichts und können insbesondere Interferenzeffekte beseitigen. Außerdem betrachten wir das Rendern von Schatten in Echtzeitanwendungen. Speichert man Momente in einer Shadow Map, kann man diese direkt filtern und so effizient Aliasing vorbeugen. Durch diese direkte Filterung wird auch die Darstellung von weichen Schatten und atmosphärischer Lichtstreuung ermöglicht.

Kennen wir zwei Potenzmomente einer Verteilung auf der reellen Achse, so ergibt sich daraus der Erwartungswert und die Varianz, die angibt, wie sehr die Verteilung davon abweicht, all ihre Masse im Erwartungswert zu lokalisieren. Weit weniger bekannt sind die Verallgemeinerungen dieses Resultats auf mehr Momente. Kennen wir $2 \cdot n$ Momente der Verteilung, so können wir $n$ Punkte und $n$ zugehörige Massen berechnen und angeben wie sehr die Verteilung davon abweicht, all ihre Masse in ebendiesen $n$ Punkten zu lokalisieren.

In der vorliegenden Dissertation [Pe17a] entwickeln wir aus dieser klassischen Theorie eine Vielzahl von effizienten und numerisch stabilen Algorithmen. Damit erreichen wir außerordentlich schnelle transiente Bildgebung, d.h. wir können für jeden Pixel einer speziellen Kamera rekonstruieren wie viel Licht mit einer bestimmten Laufzeit dorthin zurückgekehrt ist [Pe15]. Außerdem speichern wir Momente in Shadow Maps um Schatten in Echtzeit zu rendern. Solche Moment Shadow Maps können direkt gefiltert werden. So vermeiden wir Aliasing [PK15], rendern Schatten für transparente Objekte, generieren weiche Schatten und berechnen atmosphärische Lichtstreuung [Pe17b]. All diese Techniken haben geringe Kosten pro Pixel auf dem Bildschirm und eignen sich entsprechend gut für 4k Auflösungen und virtuelle Realität.

[1] Englischer Titel der Dissertation: “Moment-Based Methods for Real-Time Shadows and Fast Transient Imaging”

[2] Karlsruher Institut für Technologie, christoph.peters@kit.edu

# 1 Transiente Bildgebung

Die Lichtlaufzeitmessung ermöglicht es Tiefenbilder aufzunehmen. Eine Kamera, zusammen mit einer aktiven Beleuchtungseinheit, misst wie lange ein Lichtpuls braucht bis er von der Szene zu einem Pixel der Kamera zurückgeworfen wird. Dabei ist Mehrwegempfang ein gängiges Problem. Durch globale Beleuchtungseffekte erreicht das Licht den Pixel auf Pfaden unterschiedlicher Länge. Somit gibt es keine eindeutige Lichtlaufzeit. Vielmehr muss man von einer beliebigen, zeitlich aufgelösten Impulsantwort $G : \mathbb{R} \to [0, \infty)$ ausgehen.

Kennen wir diese Impulsantwort für jeden Pixel, so sprechen wir von einem transienten Bild. Ihre direkte Aufnahme ist in jeder Hinsicht mit hohen Kosten verbunden [Ve13]. Praktikabler ist eine indirekte Messung durch sogenannte Amplitude Modulated Continuous Wave (AMCW) Lidar Systeme. Durch eine Modulation der Lichthelligkeit und der Sensorempfindlichkeit wird dabei die Korrelation der Impulsantwort mit einem periodischen Signal gemessen. Mit Messungen bei sehr vielen Frequenzen [He13] oder unter Annahme eines einfachen Modells [GCD12] lässt sich daraus die Impulsantwort rekonstruieren.

## 1.1 Rekonstruktion von Impulsantworten

Im Folgenden beschreiben wir ein neues Rekonstruktionsverfahren, das dank einer Lösung in geschlossener Form algorithmisch effizient ist, keine Modellannahmen benötigt und bereits aus Messungen bei wenigen Frequenzen komplizierte Impulsantworten akkurat rekonstruieren kann. Dazu konfigurieren wir das AMCW Lidar System so, dass die Korrelation mit einem sinusförmigen Signal gemessen wird [Pa10].

Wir fixieren eine Grundfrequenz $f$ (z.B. $f = 23\,\mathrm{MHz}$) und betrachten die Impulsantwort fortan in Abhängigkeit von der Phase $\varphi \in (0,\ 2 \cdot \pi]$ zu dieser Grundfrequenz:

$$F(\varphi) := \sum_{l \in \mathbb{Z}} G\left(\frac{\varphi + l \cdot 2 \cdot \pi}{2 \cdot \pi \cdot f}\right)$$

Die dadurch entstehende Phasenmehrdeutigkeit kann später durch entsprechende Verfahren aufgelöst werden. Durch das AMCW Lidar System messen wir nun sogenannte trigonometrische Momente dieser Dichtefunktion:

$$c_j := \int \cos(j \cdot \varphi) \cdot F(\varphi)\,\mathrm{d}\varphi + i \cdot \int \sin(j \cdot \varphi) \cdot F(\varphi)\,\mathrm{d}\varphi = \int \exp(i \cdot j \cdot \varphi) \cdot F(\varphi)\,\mathrm{d}\varphi \in \mathbb{C}$$

Solche Messwerte benötigen wir für $j \in \{0, 1, \ldots, m\}$ wobei $m \in \mathbb{N}$. Dabei kommt $c_0 \in \mathbb{R}$ eine besondere Bedeutung zu, da es schlicht misst wie viel Licht den Pixel insgesamt erreicht. Für $j < 0$ gilt da $F$ reell ist $c_j = \overline{c_{-j}}$. Die trigonometrischen Momente sind schlicht Fourierkoeffizienten einer Verteilung.

Zur Rekonstruktion benötigen wir die Toeplitz-Matrix

$$C(c) = (c_{j-k})_{j,k=0}^{m} = \begin{pmatrix} c_0 & \overline{c_1} & \cdots & \overline{c_m} \\ c_1 & c_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \overline{c_1} \\ c_m & \cdots & c_1 & c_0 \end{pmatrix} \in \mathbb{C}^{(m+1)\times(m+1)},$$

die momentengenerierende Funktion

$$\mathbf{c}(\varphi) := (1, \exp(i \cdot 1 \cdot \varphi), \ldots, \exp(i \cdot m \cdot \varphi))^{\mathsf{T}} \in \mathbb{C}^{m+1}$$

und den kanonischen Basisvektor $e_0 := (1, 0, \ldots, 0)^{\mathsf{T}}$. Unsere Rekonstruktion ist dann gegeben durch

$$D(\varphi) := \frac{1}{2 \cdot \pi} \cdot \frac{e_0^{\mathsf{T}} \cdot C^{-1}(c) \cdot e_0}{|e_0^{\mathsf{T}} \cdot C^{-1}(c) \cdot \mathbf{c}(\varphi)|^2} \in \mathbb{R}. \tag{1}$$

Dieses Verfahren ist bekannt als Maximum Entropy Spectral Estimate [Bu75] und hat bemerkenswerte Eigenschaften.

**Theorem 1.** *Angenommen $C(c)$ ist positiv definit. Dann ist die oben definierte Dichtefunktion $D$ positiv, realisiert exakt die trigonometrischen Momente $c_0, \ldots, c_m$ und minimiert unter allen solchen Funktionen die Burg-Entropie*

$$\int_0^{2\cdot\pi} -\log D(\varphi)\, \mathrm{d}\varphi.$$

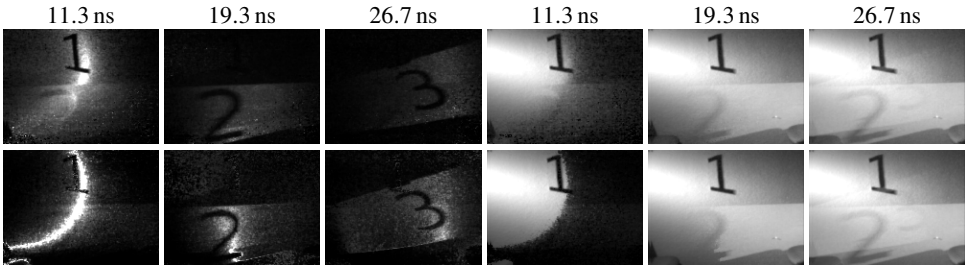
*Ist $C(c)$ nicht positiv definit, so existiert keine solche Funktion.*

Eine exakte Reproduktion der gegebenen trigonometrischen Momente ließe sich auch durch eine einfache Fourier-Reihe erreichen. Eine solche lineare Rekonstruktion kann aber keine Positivität garantieren. Der Raum der positiven Verteilungen zu gegebenen trigonometrischen Momenten ist in vielen relevanten Fällen erstaunlich klein. Besteht die ursprüngliche Verteilung aus $m$ oder weniger Dirac-$\delta$ Stößen, so enthält der Raum nur eine einzige Verteilung. In diesem Fall lässt sich mit dem sogenannten Pisarenko-Schätzer, welcher den Grenzfall von Gleichung (1) darstellt, eine perfekte Rekonstruktion erhalten. Ansonsten liefert Gleichung (1) eine Rekonstruktion mit $m$ oder weniger lokalen Maxima. Die Burg-Entropie führt dabei als Prior zu einem wohldefinierten und glatten Ergebnis, das Masse nur dann stark lokalisiert, wenn die Daten dies erzwingen.

### 1.2 Anwendungen

Der Fall einer Impulsantwort mit Dirac-$\delta$ Stößen ist in der Praxis äußerst relevant. Die Impulsantwort über den direkten Pfad entspricht stets einem Dirac-$\delta$ Stoß und Gleiches gilt

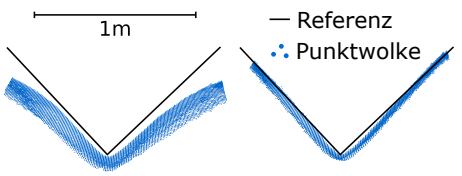
für Beiträge durch stark spekulare Flächen, z.B. Spiegel oder Glasscheiben. Wir haben unser Verfahren mit Daten aus einem modifizierten PhotonICs 19k-S3 von pmdTechnologies getestet. Abbildung 1a zeigt das rekonstruierte transiente Bild zu einer Szene mit zwei Spiegeln. Selbst mit $m = 3$ positiven Frequenzen werden bereits bis zu drei Impulse auf einem einzelnen Pixel korrekt rekonstruiert. Mit mehr Frequenzen und längerer Belichtungszeit erhält man eine schärfere Rekonstruktion mit weniger Rauschen. Überdies können wir effizient die Verteilungsfunktion rekonstruieren, d.h. errechnen wie viel Licht bis zu einem gewissen Zeitpunkt zurückgekehrt ist (Abbildung 1b).



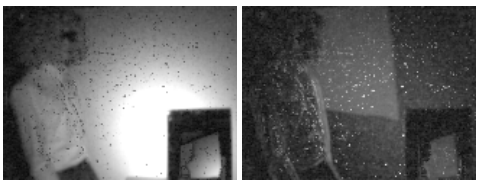

(a) Die Dichtefunktion zu drei Zeitpunkten. (b) Die Verteilungsfunktion zu drei Zeitpunkten.

Abb. 1: Transiente Bilder zu einer Szene mit zwei Spiegeln. Der direkte Puls läuft von links nach rechts, ein Spiegel reflektiert ihn nach links, ein weiterer Spiegel reflektiert ihn nach rechts. Für die obere Zeile betrug die Messzeit 91 ms bei einer Grundfrequenz von $f = 23\,\text{MHz}$ und $m = 3$ positiven Frequenzen. Unten wurde 8,21 s mit $f = 11.5\,\text{MHz}$ und $m = 8$ gemessen.

Es ist auch möglich in geschlossener Form die lokalen Maxima der rekonstruierten Dichte zu errechnen. Diese sind Kandidaten für die Lichtlaufzeit der direkten Impulsantwort. Auf diese Weise lassen sich dann Verzerrungen durch Mehrwegempfang entfernen (Abbildung 2a). Unser Ansatz ist dabei zur Aufnahme bei interaktiven Frameraten geeignet. Mit unserem Hardwareprototypen konnten wir bis zu 18.6 transiente Bilder pro Sekunde aufnehmen. Abbildung 2b zeigt das Ergebnis der Auftrennung eines solchen Bildes in direkte und indirekte Beleuchtung.




(a) Tiefenmessung bei einer Frequenz (links) und mit unserer Beseitigung des Mehrwegempfangs (rechts).



(b) Direkte Beleuchtung (links) und indirekte Beleuchtung (rechts)

Abb. 2: Eine Ansicht einer rechtwinkligen Ecke von oben mit rekonstruierten Punktwolken (links) und Aufspaltung der Beleuchtung für ein in 54 ms aufgenommenes transientes Bild (rechts).

## 2 Echtzeitschatten

Das Rendern von Schatten mag auf den ersten Blick ein völlig anderes Problem sein. Es profitiert aber von ganz ähnlichen Techniken. Der bei weitem gängigste Ansatz zur Darstellung von Schatten in Echtzeitanwendungen ist Shadow Mapping. Dabei wird die Szene aus der Perspektive der Lichtquelle rasterisiert. Pro Pixel wird ein Tiefenwert gespeichert, um den Beginn des Schattens entlang des Lichtstrahls zu erfassen.

Dieser bildbasierte Ansatz skaliert hervorragend auf komplexe Szenen, krankt aber an starkem Aliasing durch die Abtastung in der Shadow Map und auf dem Bildschirm. Percentage-closer Filtering [RSC87] beseitigt dies, indem es für jeden einzelnen Pixel eine Umgebung in der Shadow Map abtastet und die Ergebnisse des Schattentests filtert. Diese Technik ist weit verbreitet, wird aber sehr kostspielig wenn man die Filtergröße so wählt, dass Aliasing effektiv bekämpft wird. Es ist bekannt, dass man die Shadow Map direkt filtern kann, wenn man Potenzmomente der Tiefenverteilung speichert [DL06]. Vorhandene Ansätze gehen aber nicht über das zweite Moment hinaus und entsprechend ungenau ist die Rekonstruktion der Schatten.

### 2.1 Moment Shadow Mapping

Unsere Moment Shadow Map [PK15] ist eine Textur mit $m = 4$ Kanälen. Der erste speichert die Tiefe $z \in [-1, 1]$, die in einer gewöhnlichen Shadow Map gespeichert werden würde. Die weiteren drei Kanäle speichern die Potenzen $z^2$, $z^3$ und $z^4$. Wenden wir einen Filter mit $n \in \mathbb{N}$ Gewichten $w_0, \ldots, w_{n-1} \geq 0$ an, der Pixel mit Tiefen $z_0, \ldots, z_{n-1} \in [-1, 1]$ kombiniert, so erhalten wir Potenzmomente einer Tiefenverteilung $Z$:

$$Z := \sum_{l=0}^{n-1} w_l \cdot \delta_{z_l}, \qquad b_j := \mathcal{E}_Z(\mathbf{z}^j) = \sum_{l=0}^{n-1} w_l \cdot z_l^j \text{ für } j \in \{0, \ldots, m\}. \tag{2}$$

Das Ziel beim Filtern von Schatten ist es zu berechnen, welcher gewichtete Anteil der abgetasteten Tiefen zum Schatten beiträgt:

$$Z(\mathbf{z} < z_f) = \sum_{l=0}^{n-1} w_l \cdot \begin{cases} 1 & \text{falls } z_l < z_f, \\ 0 & \text{sonst.} \end{cases}$$

Algorithmus 1 berechnet anhand der Momente $b_0, \ldots, b_m$ die bestmögliche untere Schranke zu dieser Größe. Durch die systematische Unterschätzung wird vermieden, dass Flächen fälschlicher Weise einen Schatten auf sich selbst werfen.

Die Eingabe für Algorithmus 1 ist gültig wenn, $B(b)$ positiv definit ist. Dann ist auch der Algorithmus erfolgreich. Erneut ist der Grenzfall interessant. Wenn $Z$ sich aus nur $\frac{m}{2}$ Punkten zusammensetzt, so ist es eindeutig durch die Momente $b_0, \ldots, b_m$ bestimmt und

**Algorithmus 1** Berechnung der Schattenintensität mit Moment Shadow Mapping.
**Eingabe:** Oberflächentiefe $z_f \in \mathbb{R}$ und Potenzmomente $b_0, \ldots, b_m \in \mathbb{R}$ für gerades $m$.
**Ausgabe:** Eine scharfe untere Schranke zu $Z(\mathbf{z} < z_f)$.

- Setze $B(b) := (b_{j+k})_{j,k=0}^{\frac{m}{2}} = \begin{pmatrix} b_0 & b_1 & \cdots & b_{\frac{m}{2}} \\ b_1 & b_2 & \iddots & b_{\frac{m}{2}+1} \\ \vdots & \iddots & \iddots & \vdots \\ b_{\frac{m}{2}} & b_{\frac{m}{2}+1} & \cdots & b_m \end{pmatrix} \in \mathbb{R}^{(\frac{m}{2}+1)\times(\frac{m}{2}+1)}$.
- Löse $B(b) \cdot q = (1, z_f^1, \ldots, z_f^{\frac{m}{2}})^\mathsf{T}$ per Cholesky-Zerlegung nach $q \in \mathbb{R}^{\frac{m}{2}+1}$.
- Löse $\sum_{j=0}^{\frac{m}{2}} q_j \cdot z^j = 0$ nach $z$ um Nullstellen $z_1, \ldots, z_{\frac{m}{2}} \in \mathbb{R}$ zu erhalten.
- Setze $z_0 = z_f$ und für $l \in \{0, \ldots, \frac{m}{2}\}$ setze $v_l := 1$ wenn $z_l < z_0$ und $v_l := 0$ sonst.
- Berechne den Rückgabewert

$$\begin{pmatrix} b_0 \\ \vdots \\ b_{\frac{m}{2}} \end{pmatrix}^\mathsf{T} \cdot \begin{pmatrix} 1 & z_0^1 & \cdots & z_0^{\frac{m}{2}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{\frac{m}{2}}^1 & \cdots & z_{\frac{m}{2}}^{\frac{m}{2}} \end{pmatrix}^{-1} \cdot \begin{pmatrix} v_0 \\ \vdots \\ v_{\frac{m}{2}} \end{pmatrix}.$$

wird perfekt rekonstruiert. Die relativ kleinen Filterbereiche für die Schatten, enthalten selten mehr als zwei Flächen. Daher erhält man mit $m = 4$ bereits hervorragende Rekonstruktionen.

In der Praxis bringt dieser Fall aber auch Herausforderungen mit sich. Er korrespondiert zu einer positiv semi-definiten aber singulären Matrix $B(b)$. Daher haben wir bei der Implementierung sorgfältig auf numerische Stabilität geachtet. Gleichzeitig muss die Implementierung auf den bei GPUs üblichen SIMD-Architekturen sehr effizient sein. Unsere Implementierung für $m = 4$ benötigt nur 32 Fließkommaoperationen und keine Zweige.

Ein weiteres Problem sind Rundungsfehler in der Eingabe. Es ist wichtig den Bedarf an Speicher und Bandbreite minimal zu halten. Wir möchten nur 16 bit pro Moment aufbringen. Hierzu wenden wir eine affine Abbildung auf die Momente an bevor wir sie speichern. Diese ist dazu optimiert die Entropie der Daten zu maximieren ohne Werte außerhalb des darstellbaren Zahlenbereichs zu generieren. Damit die verbleibenden Rundungsfehler nicht zu ungültigen Eingaben führen, wird ein optimiertes Bias in der Größenordnung des Rundungsfehlers angewandt.

### 2.2 Optimalität von Potenzmomenten

Die Verwendung von Potenzmomenten bringt erhebliche Vorteile mit sich, da die in Algorithmus 1 verwendete Lösungsstrategie bekannt ist [KN77]. Dennoch stellt sich die Frage ob man bessere Ergebnisse erhalten könnte wenn man andere Daten in den vier Kanälen der Shadow Map speichert. In der Tat kann man $z_l^j$ in Gleichung (2) durch eine beliebige, stetige Funktion $\mathbf{a}_j : [-1, 1] \to \mathbb{R}$ ersetzen und erhält immernoch eine wohldefinierte scharfe, untere Schranke. Diskretisiert man das Intervall $[-1, 1]$, so lässt sich diese sogar effizient berechnen. Die Zielfunktion $Z(\mathbf{z} < z_f)$ ist linear und wir haben lineare Gleichungen und Ungleichungen als Nebenbedingungen. Es handelt sich also um ein Problem der linearen Programmierung. Zum Rendering ist dieser Ansatz zu langsam, doch er ermöglicht interessante Einblicke.

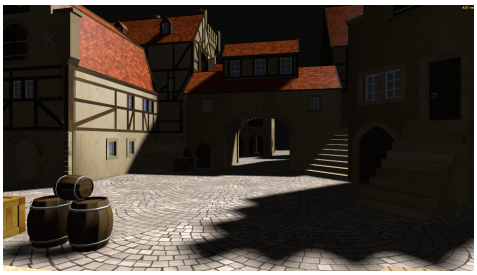
Von dieser Erkenntnis ausgehend, haben wir Testfälle definiert in denen auf mehreren Szenen gefilterte Schatten errechnet werden. Als Kandidaten haben wir 66045 unterschiedliche Kombinationen von Funktion $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4$ betrachtet. Die Evaluation lief mehrere Wochen auf einem CPU-Cluster. Die Ergebnisse zeigen, dass über 15000 Kandidaten nahezu identisch gute Ergebnisse liefern. Einen besseren Kandidaten gibt es nicht. Erfreulicherweise gehört Moment Shadow Mapping zu diesen Kandidaten. Der Grund dürften die theoretisch günstigen Eigenschaften für Tiefenverteilungen mit nur zwei Tiefen sein. Dieses Verhalten ist in der Praxis äußerst wichtig und auch andere Kandidaten legen es an den Tag.

### 2.3 Anwendungen von Moment Shadow Maps

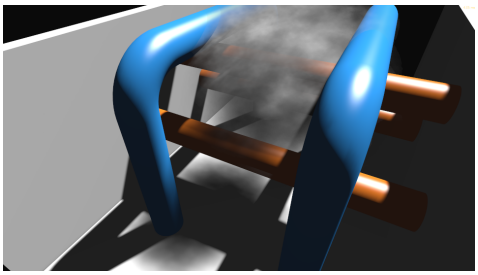
Dass man Moment Shadow Maps direkt filtern kann, ermöglicht vielfältige Anwendungen. Die naheliegendste ist Antialiasing für harte Schatten. Dabei generiert man die Moment Shadow Map mit hardwarebeschleunigtem Multisample Antialiasing und wendet dann zusätzlich in zwei Durchläufen einen gaußschen Weichzeichner an (Abbildung 3a). Beim Generieren der Moment Shadow Map kann man auch transparente Schattenwerfer durch schlichtes Alpha Blending rendern (Abbildung 3b).

Ist die Lichtquelle kein Punktlicht sondern ein Flächenlicht, so ist der Schatten am Kontaktpunkt von Schattenwerfer und -empfänger hart und wird dann zunehmend weicher. Dieser Effekt lässt sich durch gefilterte harte Schatten mit variabler Filtergröße annähern [Fe05]. Wir generieren eine Summed Area Table [Cr84] für eine Moment Shadow Map und können so in konstanter Zeit schätzen wie groß der Filter sein muss und entsprechend Filtern (Abbildung 3c).

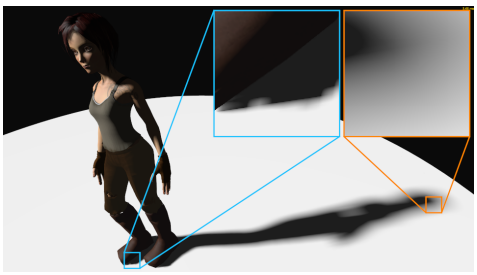
Schließlich ist es auch interessant Schatten in streuenden Medien wie etwa Wasserdampf oder staubiger Luft zu betrachten. Hierbei besteht die Herausforderung darin, dass man den Schatten entlang des Sichtstrahls aufintegrieren muss. Das Integral pro Pixel durch numerische Quadratur anzunähern ist aber kostspielig. Wir überführen stattdessen eine



(a) Moment Shadow Mapping zum Antialiasing harter Schatten (4.0 ms)



(b) Moment Shadow Mapping für transparente Schattenwerfer (4.1 ms)



(c) Moment soft Shadow Mapping (3.5 ms)

(d) Vorgefilterte Lichtstreuung mit sechs Potenzmomenten (5.4 ms)

Abb. 3: Die Anwendung von Moment Shadow Maps auf unterschiedliche Lichttransportphänomene. Zeiten beziehen sich auf das vollständige Rendern eines Bildes auf einer NVIDIA GeForce GTX 970.

Moment Shadow Map mit $m = 6$ Momenten in ein Koordinatensystem in dem Sichtstrahlen zu Texturzeilen korrespondieren. Dann generieren wir gewichtete Prefixsummen über die Zeilen und berechnen so die Lichtstreuung für jeden möglichen Sichtstrahl vor (Abbildung 3d).

All diese Techniken haben vergleichsweise hohe Kosten pro Texel der Moment Shadow Map, gelangen dann aber pro Pixel auf dem Bildschirm mit wenigen Texturzugriffen und arithmetischen Operationen zum Ziel. Sie skalieren daher hervorragend auf hohe Auflösungen. Abbildung 3 benutzt z.B. eine Auflösung von 3840×2160 für alle Bilder und schließt das Rendern dennoch in wenigen Millisekunden ab.

## 3 Fazit

Die Theorie der Momente ist ein mächtiges Werkzeug, genießt in der Informatik aber keine große Bekanntheit. Momente zu errechnen ist eine trivial parallelisierbare Operation. Die

Momente sind kompakt und lassen sich leicht mit neuen Daten kombinieren. Die Rekonstruktionen sind robust und erreichen eine enorme Genauigkeit wenn die ursprüngliche Verteilung ihre Masse um wenige Punkte lokalisiert. Dadurch erhält man gewissermaßen ein eindimensionales Clustering der Daten. Compressed Sensing Ansätze haben ähnliche Eigenschaften, sind aber mit erheblich größerem Rechenaufwand verbunden. Auch jenseits der Computergraphik dürften die von uns entwickelten effizienten und robusten Implementierungen von Interesse sein.

Mit Blick auf die bisher erforschten Anwendungen ist ein großes industrielles Interesse erkennbar. Es sind bereits mehrere Spiele auf dem Markt, die standardmäßig Moment Shadow Mapping nutzen[3]. Auch die transiente Bildgebung und die damit verbundene Beseitigung des Mehrwegempfangs hat bei den Entwicklern entsprechender Kameras großes Interesse geweckt.

Der nächste große Schritt wird sein, vergleichbare Algorithmen für höherdimensionale Probleme zu entwickeln. Zweidimensionale Momentenprobleme sind gegenwärtig ein lebendiges Forschungsfeld [La11]. Dabei stößt man deutlich schneller an Grenzen als bei den eindimensionalen Problemen. Dennoch bergen die Arbeiten großes Potenzial. So könnten z.B. momentenbasierte Rekonstruktionen auf der Sphäre im $\mathbb{R}^3$ immensen Nutzen für die Computergraphik entfalten.

## Literaturverzeichnis

[Bu75] Burg, John Parker: Maximum Entropy Spectral Analysis. Ph.D. dissertation, Stanford University, Department of Geophysics, 1975.

[Cr84] Crow, Franklin C.: Summed-Area Tables for Texture Mapping. In: Proceedings of the 11th annual conference on Computer graphics and interactive techniques. SIGGRAPH '84. ACM, S. 207–212, 1984.

[DL06] Donnelly, William; Lauritzen, Andrew: Variance Shadow Maps. In: Proceedings of the 2006 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games. i3D '06. ACM, S. 161–165, 2006.

[Fe05] Fernando, Randima: Percentage-Closer Soft Shadows. In: ACM SIGGRAPH 2005 Sketches. ACM, 2005.

[GCD12] Godbaz, John P.; Cree, Michael J.; Dorrington, Adrian A.: Closed-form inverses for the mixed pixel/multipath interference problem in AMCW lidar. Proc. SPIE, 8296:829618-1–829618-15, 2012.

[He13] Heide, Felix; Hullin, Matthias B.; Gregson, James; Heidrich, Wolfgang: Low-budget Transient Imaging Using Photonic Mixer Devices. ACM Trans. Graph. (Proc. SIGGRAPH 2013), 32(4):45:1–45:10, July 2013.

[3] Nach Aussage der Entwickler gilt dies für die Titel Deformers, Lone Echo und Echo Arena der Ready at Dawn Studios sowie Injustice 2 der Netherrealm Studios. Zusätzlich ist mit einer Dunkelziffer zu rechnen.

[KN77] Kreĭn, Mark Grigorievich; Nudel'man, Adol'f Abramovich: The Markov Moment Problem and Extremal Problems, Jgg. 50 in Translations of Mathematical Monographs. American Mathematical Society, 1977.

[La11] Lasserre, Jean Bernard: Moments, Positive Polynomials and Their Applications, Jgg. 1 in Series on Optimization and Its Applications. Imperial College Press, 2011.

[Pa10] Payne, Andrew D.; Dorrington, Adrian A.; Cree, Michael J.; Carnegie, Dale A.: Improved measurement linearity and precision for AMCW time-of-flight range imaging cameras. Appl. Opt., 49(23):4392–4403, August 2010.

[Pe15] Peters, Christoph; Klein, Jonathan; Hullin, Matthias B.; Klein, Reinhard: Solving Trigonometric Moment Problems for Fast Transient Imaging. ACM Trans. Graph. (Proc. SIGGRAPH Asia 2015), 34(6), November 2015.

[Pe17a] Peters, Christoph: Moment-Based Methods for Real-Time Shadows and Fast Transient Imaging. Dissertation, University of Bonn, Dezember 2017. urn:nbn:de:hbz:5n-49187.

[Pe17b] Peters, Christoph; Münstermann, Cedrick; Wetzstein, Nico; Klein, Reinhard: Improved Moment Shadow Maps for Translucent Occluders, Soft Shadows and Single Scattering. Journal of Computer Graphics Techniques (JCGT), 6(1):17–67, März 2017.

[PK15] Peters, Christoph; Klein, Reinhard: Moment Shadow Mapping. In: Proceedings of the 19th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games. i3D '15. ACM, S. 7–14, März 2015.

[RSC87] Reeves, William T.; Salesin, David H.; Cook, Robert L.: Rendering Antialiased Shadows with Depth Maps. In: Proceedings of the 14th annual conference on Computer graphics and interactive techniques. SIGGRAPH '87. ACM, S. 283–291, 1987.

[Ve13] Velten, Andreas; Wu, Di; Jarabo, Adrian; Masia, Belen; Barsi, Christopher; Joshi, Chinmaya; Lawson, Everett; Bawendi, Moungi; Gutierrez, Diego; Raskar, Ramesh: Femto-photography: Capturing and Visualizing the Propagation of Light. ACM Trans. Graph. (Proc. SIGGRAPH 2013), 32(4):44:1–44:8, Juli 2013.

**Christoph Peters** ist seit Januar 2018 Post-Doktorand am Karlsruher Institut für Technologie in der Gruppe von Carsten Dachsbacher. Zuvor hat er ein halbes Jahr bei NVIDIA in Redmond, WA zum Echtzeitrendering geforscht. Seine Promotion hat er im Mai 2017 in der Gruppe von Reinhard Klein an der Universität Bonn mit Auszeichnung abgeschlossen. Davor erhielt er 2013 in Bonn den M.Sc. in Informatik und 2011 in Köln den B.Sc. in Mathematik. Schwerpunkt seiner Arbeit ist die Entwicklung neuer Renderingverfahren. Dabei durchforstet er die mathematische Literatur nach Lösungsansätzen, die sich effizient auf massiv parallelen Prozessoren implementieren lassen und praxistaugliche Techniken für relevante Probleme liefern.

# Neue Perspektiven auf die Kostenpartitionierung für optimale klassische Handlungsplanung[1]

Florian Pommerening[2]

**Abstract:** Zulässige Heuristiken sind die wichtigste Zutat beim Lösen von klassischen Handlungsplanungsproblemen mit heuristischer Suche. Da größere zulässige Heuristikwerte genauer sind, untersuchen wir, wie man Heuristikwerte so miteinander kombinieren kann, dass die Kombination zulässig, aber besser als das Maximum ist. Die Dissertation enthält drei neue Beiträge dazu. Mit Erweiterungen der bekannten Technik der *Kostenpartitionierung* können größere Heuristikwerte aus den gleichen Heuristiken gewonnen werden. Die neue Familie der *Operator-Counting-Heuristiken* vereint viele existierende Heuristiken und erlaubt sie auf eine neue Art zu kombinieren. Schließlich kann mit der neuen Familie der *Potentialheuristiken* die Suche nach guten Heuristiken selbst als Optimierungsproblem aufgefasst werden. Die beiden neuen Heuristikfamilien sind eng mit der Kostenpartitionierung verwandt. Sie bieten eine neue Perspektive auf kostenpartitionierte Heuristiken und haben schon jetzt Forschung in und außerhalb der klassischen Handlungsplanung angeregt.

## 1 Einführung

Das Ziel der Handlungsplanung [GNT04] ist es, Aktionssequenzen zu finden, die ein gegebenes Ziel erfüllen. In einem Logistikproblem sind beispielsweise die Aktionen das Ein- und Ausladen von Paketen in Lieferwagen und die Bewegungen der Fahrzeuge zwischen verschiedenen Orten. In einer gegebenen Situation wird dann ein Plan gesucht, mit dem alle Pakete mit möglichst wenigen Aktionen ausgeliefert werden können. Weitere Beispiele kommen aus verschiedensten Gebieten, von der Missionsplanung für Mars-Rover über industrielle Anwendungen bis hin zu kombinatorischen Spielen wie dem Schiebepuzzle.

Die Dissertation [Po17] legt den Fokus auf *klassische* Handlungsplanung, in der alle Aktionen diskret und deterministisch sind, alle Effekte vollständig beobachtbar sind und sequenzielle Pläne für einzelne Agenten vor der Ausführung gesucht werden. In der *optimalen* Handlungsplanung müssen insbesondere die Kosten der gefundenen Pläne minimal sein. Um optimale Pläne zu finden, wird oft die A$^*$-Suche [HNR68] verwendet. A$^*$ wählt in jedem Schritt einen vorher erreichten Zustand und *expandiert* diesen. Dabei wird für jede in diesem Zustand anwendbare Aktion der eindeutige Nachfolgezustand generiert. Die Auswahl des zu expandierenden Zustands $s$ basiert auf der Summe der Kosten, um $s$ zu erreichen, und einer Schätzung der Kosten, um von $s$ das nächstgelegene Ziel zu erreichen. Letztere wird dabei von einer *Heuristik* [Pe84] berechnet. A$^*$ garantiert eine optimale Lösung, wenn die verwendete Heuristik *zulässig* ist, das heißt, wenn sie die tatsächlichen Kosten nie überschätzt. Üblicherweise führt eine zulässige Heuristik mit höheren Werten schneller zum Ziel, da die Schätzungen genauer werden.

[1] Englischer Titel der Dissertation: "New Perspectives on Cost Partitioning for Optimal Classical Planning"

[2] Departement Mathematik und Informatik, Universität Basel, florian.pommerening@unibas.ch

Es gibt viele zulässige Heuristiken, und oft hängt es von der Anwendung, der Probleminstanz, oder sogar vom Zustand ab, welche die besten Werte liefert. Daher werden oft mehrere Heuristiken verwendet. Das wirft die Frage auf, mit der wir uns hier beschäftigen: Wie können mehrere zulässige Heuristiken auf zulässige Art kombiniert werden?

Es gibt zwei etablierte Antworten auf diese Frage. Das *Maximum* von zulässigen Heuristiken ist immer zulässig und mindestens so gut wie jede einzelne Heuristik. Mit der *Kostenpartitionierung* [KD10] wird jede Heuristik auf einer Kopie des Problems ausgewertet, die sich nur in den Aktionskosten unterscheidet. Die ursprünglichen Kosten werden dabei so auf die Kopien verteilt, dass die Summe der Heuristikwerte zulässig bleibt. Die optimale Aufteilung der Kosten kann für bestimmte Heuristiken effizient berechnet werden und führt zu Werten, die mindestens so gut wie das Maximum sind (und oft deutlich besser).

Wir erweitern die Kostenpartitionierung, so dass die gleichen Heuristiken zu höheren zulässigen Werten kombiniert werden können. Wir führen außerdem zwei neue Heuristikfamilien ein, die neue Perspektiven auf die Kostenpartitionierung bieten, noch höhere Werte erreichen können und dabei helfen, bekannte Heuristiken besser zu verstehen. Vorher definieren wir noch die dafür benötigte Notation aus der Handlungsplanung.

## 2 Handlungsplanung

Wir betrachten in dieser Arbeit Handlungsplanungsprobleme in SAS$^+$ [BN95]. Ein solches Problem ist ein Tupel $\langle \mathcal{V}, \mathcal{O}, s_\mathrm{I}, s_\star, \mathit{cost} \rangle$ mit den folgenden Komponenten: $\mathcal{V}$ ist eine endliche Menge von *Variablen*, wobei jede Variable $V \in \mathcal{V}$ einen endlichen Wertebereich $\mathit{dom}(V)$ besitzt. Eine *partielle Variablenbelegung* ist eine Funktion $p$, die einen Teil der Variablen $\mathit{vars}(p) \subseteq \mathcal{V}$ auf Werte in ihren Wertebereichen abbildet. Wenn $\mathit{vars}(p) = \mathcal{V}$, sprechen wir von einem *Zustand*. Der *Startzustand* $s_\mathrm{I}$ ist ein Zustand, und das *Ziel* $s_\star$ ist eine partielle Variablenbelegung. Wenn ein Zustand $s$ in allen Variablen mit einer partiellen Variablenbelegung $p$ übereinstimmt ($s(V) = p(V)$ für alle $V \in \mathit{vars}(p)$) sagen wir $s$ ist konsistent mit $p$. Zustände, die konsistent mit dem Ziel sind, nennen wir *Zielzustände*.

Die Menge $\mathcal{O}$ enthält endlich viele *Operatoren* $o = \langle p, e \rangle$ wobei sowohl die *Vorbedingung* $\mathit{pre}(o) = p$ als auch der *Effekt* $\mathit{eff}(o) = e$ partielle Variablenbelegung sind. Ein Operator $o$ ist in einem Zustand $s$ anwendbar, wenn $s$ konsistent mit $\mathit{pre}(o)$ ist. In diesem Fall ist das Ergebnis der Anwendung der Zustand $s[\![o]\!]$ mit $s[\![o]\!](V) = \mathit{eff}(o)(V)$ für alle $V \in \mathit{vars}(\mathit{eff}(o))$ und $s[\![o]\!](V) = s(V)$ für alle anderen Variablen. Ein *$s$-Plan* ist eine endliche Folge $\langle o_1, \ldots, o_n \rangle$ von Operatoren, die nacheinander im Zustand $s$ anwendbar sind, so dass $s[\![o_1]\!][\![\ldots]\!][\![o_n]\!]$ ein Zielzustand ist. Einen $s_\mathrm{I}$-Plan nennen wir einfach nur *Plan*. Die *Kostenfunktion* $\mathit{cost} : \mathcal{O} \to \mathbb{R}_0$ bildet jeden Operator auf seine Kosten ab. Die Kosten einer Folge $\langle o_1, \ldots, o_n \rangle$ sind $\sum_{i=1}^{n} \mathit{cost}(o_i)$. Das Ziel der optimalen klassischen Handlungsplanung ist es, einen Plan mit minimalen Kosten, einen *optimalen Plan*, zu finden. Die Kosten eines optimalen $s$-Plans bezeichnen wir mit $h^*(s)$.

Um optimale Pläne zu finden, wird häufig A$^*$ mit einer *zulässigen Heuristik* verwendet [HNR68]. Eine Heuristik $h$ bildet Zustände auf Werte in $\mathbb{R}_0 \cup \{\infty\}$ ab, die Schätzungen für die Kosten eines $s$-Plans repräsentieren. Wenn die Heuristik nie überschätzt, wenn also

$h \leq h^*$, dann nennen wir die Heuristik *zulässig* und die von A$^*$ gefundene Lösung ist garantiert optimal. Es gibt viele Beispiele für Heuristiken in der Literatur. Die meisten lösen eine relaxierte Form des Planungsproblems und verwenden dann die optimalen Kosten des relaxierten Problems als Heuristik für das ursprüngliche Problem. Für diese Zusammenfassung beschränken wir uns auf *Abstraktionsheuristiken* [HHH08].

Eine *Abstraktionsfunktion* $\alpha$ bildet die Zustände eines Planungsproblems $\Pi$ auf eine (üblicherweise kleinere) Menge von *abstrakten Zuständen* ab. Dadurch wird ein abstraktes Planungsproblem induziert, das den Startzustand $\alpha(s_\mathrm{I})$ hat, die Zielzustände $\alpha(s_G)$ für alle Zielzustände $s_G$ von $\Pi$, und in dem eine Transition von $\alpha(s)$ nach $\alpha(s')$ mit Operator $o$ möglich ist, wenn in $\Pi$ der Operator $o$ in $s$ anwendbar ist und $s[\![o]\!] = s'$. Jeder Plan von $\Pi$ ist auch im abstrakten Problem noch anwendbar und führt auch dort mit den gleichen Kosten zu einem Zielzustand. Die Kosten eines optimalen Plans im abstrakten Problem können daher nicht höher sein als die optimalen Kosten von $\Pi$. Die Heuristikfunktion $h^\alpha$, die jeden Zustand $s$ auf die optimalen Kosten eines $\alpha(s)$-Plans im von $\alpha$ induzierten Problem abbildet, ist daher eine zulässige Heuristik ($h^\alpha \leq h^*$).

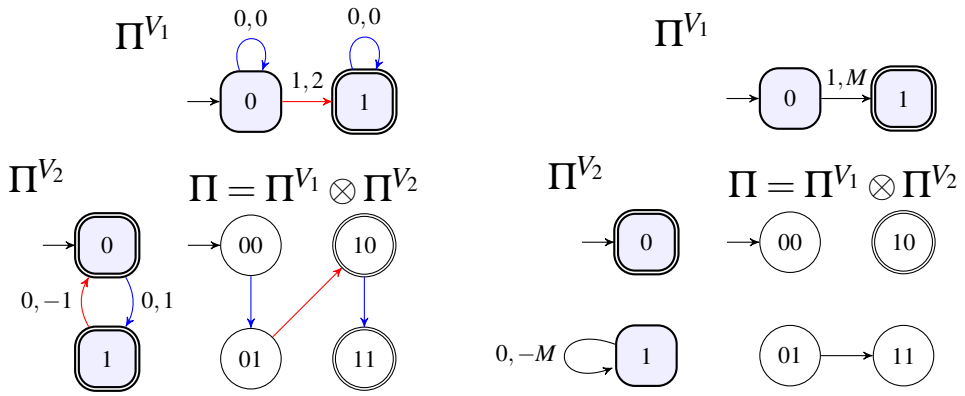
Ein wichtiger Spezialfall von Abstraktionen sind Projektionen, bei denen ein Zustand auf eine Teilmenge der Variablen projiziert wird. Die Teilmenge der Variablen wird *Pattern* genannt und die Abstraktionsheuristik für eine Projektion wird *Pattern-Database-Heuristik* (PDB-Heuristik) genannt [CS98, Ed01].

## 3 Kostenpartitionierung

Einzelne Heuristiken haben oft Schwächen in bestimmten Bereichen, was durch die Verwendung mehrerer Heuristiken ausgeglichen werden kann. Für optimales Planen stellt sich die Frage, wie diese Heuristiken kombiniert werden können, ohne deren Zulässigkeit zu verlieren. Das Maximum von mehreren zulässigen Heuristiken ist immer zulässig, aber höchstens so gut wie die beste einzelne Heuristik. Die Summe der Heuristikwerte liefert höhere Werte, ist aber nicht immer zulässig. Eine Menge von zulässigen Heuristiken heißt *additiv*, wenn auch ihre Summe zulässig ist.

Zulässige Heuristiken können additiv gemacht werden, indem sie unter angepassten Kostenfunktionen ausgewertet werden. Haslum et al. [HBG05] partitionieren die Operatoren in Mengen $\mathcal{O}_1, \ldots, \mathcal{O}_n$ und definieren $h_{\mathcal{O}_i}$ als die Heuristik $h$, die aber bei ihrer Berechnung die Kosten aller Operatoren außerhalb von $\mathcal{O}_i$ ignoriert. Die Heuristiken $h_{\mathcal{O}_1}, \ldots h_{\mathcal{O}_n}$ sind dann additiv, da die Kosten jedes Operators nur in einer der Heuristiken gezählt werden. *Kostenpartitionierung* [KD10] generalisiert diese Idee weiter. Statt die Kosten jedes Operators voll in einer Heuristik zu zählen und in allen anderen Heuristiken zu ignorieren, wird die Kostenfunktion *cost* in Funktionen $cost_i : \mathcal{O} \to \mathbb{R}_0^+$ partitioniert. Solange die ursprünglichen Kosten nicht überschritten werden ($\sum_{i=1}^n cost_i(o) \leq cost(o)$ für alle $o \in \mathcal{O}$), werden zulässige Heuristiken $h_1, \ldots, h_n$ unter den Kostenfunktionen $cost_1, \ldots, cost_n$ additiv.

Katz und Domshlak [KD10] zeigen weiterhin, dass für explizite Abstraktionsheuristiken (wie zum Beispiel PDBs) die *optimale Partitionierung* der Kosten in polynomieller Zeit mit einem linearen Programm gefunden werden kann.




(a) Projektionen auf Variablen, die nicht im Ziel vorkommen, können Informationen beitragen.

(b) Unlösbarkeit kann erkannt werden, auch wenn beide Projektionen lösbar sind.

Abb. 1: Beispielprobleme, die die Vorteile von allgemeiner Kostenpartitionierung zeigen. Die Probleme haben jeweils zwei binäre Variablen $V_1, V_2$, der eingehende Pfeil markiert den Startzustand, doppelte Umrandungen markieren Zielzustände. Die Projektionen auf die beiden Variablen werden über ($V_1$) und links neben ($V_2$) dem konkreten Zustandsraum gezeigt. Im ursprünglichen Problem kosten alle Operatoren 1. Eine mit $a, b$ beschriftete Kante kostet $a$ unter unter einer optimalen Partitionierung ohne negative Kosten und $b$ wenn negative Kosten erlaubt sind.

## 4 Erweiterungen der Kostenpartitionierung

Wir erweitern die Idee der Kostenpartitionierung in zwei verschiedene Richtungen. Zuerst beobachten wir, dass die partitionierten Kostenfunktionen bei Katz und Domshlak keine negativen Werte erlauben. Das scheint natürlich, weil die ursprüngliche Kostenfunktion auch nicht-negativ ist. Wir zeigen jedoch, dass diese Einschränkung nicht nötig ist:

**Satz 1** (Allgemeine Kostenpartitionierung [Po15])**.** *Sei $\Pi$ ein Planungsproblem mit Operatoren $\mathcal{O}$ und Kostenfunktion cost und seien $cost_1, \ldots, cost_n$ Kostenfunktionen $cost_i : \mathcal{O} \to \mathbb{R}$ mit der Eigenschaft $\sum_{i=1}^{n} cost_i(o) \leq cost(o)$ für alle $o \in \mathcal{O}$. Dann ist $\sum_{i=1}^{n} h_i(s, cost_i)$ ein zulässiger Heuristikwert für den Zustand s, wenn alle $h_i(s, cost_i)$ zulässige Heuristikwerte für s unter der Kostenfunktion $cost_i$ sind.*

Die Änderung, auch negative Kosten zuzulassen, mag klein erscheinen, jedoch sind die Auswirkungen weitreichend. Da jede Partitionierung in nicht-negative Kosten auch eine gültige Partitionierung in allgemeine Kosten ist, ist der Heuristikwert unter einer optimalen allgemeinen Partitionierung mindestens so hoch wie der unter einer optimalen nicht-negativen. Negative Kosten bieten zusätzliche Möglichkeiten, durch die höhere Heuristikwerte erreicht werden können. Abbildung 1 zeigt zwei kleine Beispiele dafür.

Unter nicht-negativen Kosten ist der Heuristikwert einer Projektion auf Variablen, die nicht im Ziel erwähnt werden, immer 0, da jeder abstrakte Zustand ein Zielzustand ist. Solche Projektionen können also unter nicht-negativen Kosten nichts zu einer Heuristikkombination beitragen. In typischen Planungsproblemen betrifft das einen Großteil der Variablen.

Im Beispiel 1a können daher nur die Kosten des roten Operators berücksichtigt werden, und eine optimale nicht-negative Kostenpartitionierung erreicht nur einen Wert von 1. Mit negativen Kosten können die Kosten des roten Operators in der Projektion auf $V_2$ auf $-1$ gesenkt werden, um sie in der Projektion auf $V_2$ auf 2 zu heben. In der Kombination ergibt sich damit ein Heuristikwert von 2, was in diesem Beispiel der perfekte Wert ist.[3]

Das Problem in Beispiel 1b ist offensichtlich unlösbar. Allerdings ist das nicht so leicht zu erkennen, wenn nur die Projektionen betrachtet werden, da beide lösbar sind. Mit nicht-negativen Kosten ist der Wert der Projektion auf $V_2$ immer 0 und der andere Wert höchstens 1. Die Unlösbarkeit kann also ohne negative Kosten nicht erkannt werden. Wie in der Abbildung gezeigt, kann allerdings mit negativen Kosten der Heuristikwert auf $0+M$ für ein beliebig großes $M$ gehoben werden. Der (in polynomieller Zeit berechenbare) optimale Wert ist in diesem Fall als $\infty$ definiert und die Unlösbarkeit wird erkannt.

In der Praxis führen allgemeine Kosten oft zu deutlich höheren Heuristikwerten, ohne die Berechnungszeit für eine optimale Partitionierung stark zu erhöhen [Po15, Po17]. Durch zusätzlich Betrachtung von Variablen, die nicht im Ziel vorkommen, werden die Heuristikwerte weiter erhöht [Po17].

Die Dissertation betrachtet noch eine zweite Erweiterungen der Kostenpartitionierung, die wir hier nur kurz ansprechen. Wird ein Operator mehrmals in einem Plan verwendet, kann es sein, dass unterschiedliche Abstraktionen die verschiedenen Verwendungen registrieren. In diesem Fall ist es hilfreich, den Kontext, in dem der Operator verwendet wird, bei der Partitionierung mit zu betrachten. Es werden dann nicht mehr die Kosten jedes Operators partitioniert, sondern die Kosten jeder Transition [Ke16]. Da die Anzahl der Transitionen exponentiell in der Anzahl der Operatoren seien kann, ist es im Allgemeinen nicht möglich, eine optimale Partitionierung in polynomieller Zeit zu finden. Wir werden allerdings später effiziente Verfahren für einige Spezialfälle vorstellen.

Wir zeigen, dass die beiden Erweiterungen unvergleichbar sind, dass es also Probleme gibt, bei denen allgemeine Kostenfunktionen zu höheren Heuristikwerten führen als ein Partitionieren der Transitionskosten, und Probleme, bei denen das Gegenteil der Fall ist. Die beiden Erweiterungen lassen sich einfach kombinieren, indem man die Kosten von Transitionen partitioniert und dabei negative Kosten zulässt. Die optimalen Werte dieser Kombination sind mindestens so gut wie die jeder Erweiterung einzeln [Ke16].

## 5 Operator-Counting-Heuristiken

Obwohl es polynomielle Verfahren gibt, um Kosten optimal zwischen Abstraktionsheuristiken zu verteilen, ist die Berechnung in der Praxis oft so teuer, dass die gesparte Suchzeit die zusätzliche Berechnungszeit nicht ausgleichen kann. Allerdings gibt es mehrere Verfahren, die gezeigt haben, dass es sich lohnen kann, in jedem Zustand ein lineares Programm zu lösen [va07, KD09, Bo13, PRH13]. Diese Heuristiken verwenden Informationen aus unterschiedlichsten Quellen und sind nicht direkt miteinander kombinierbar.

[3] Werden die Kosten in der Projektion auf $V_2$ weiter reduziert, entsteht ein Zyklus mit negativen Kosten. Diese Projektion erhält damit einen Heuristikwert von $-\infty$. Die Summe beider Heuristikwerte bleibt zulässig.

Wir bringen diese Ansätze zusammen, indem wir eine Heuristikfamilie einführen, die alle Verfahren abdeckt und erlaubt, sie miteinander zu vergleichen und zu kombinieren [Po14].

Die grundsätzliche Idee von *Operator-Counting-Heuristiken* ist, notwendige Eigenschaften von Plänen deklarativ als Ungleichungen über Variablen auszudrücken, welche die Anzahl von Operatoranwendungen beschreiben. Für deren Definition verwenden wir die Notation $occur_o(\pi)$ für die Anzahl von Vorkommen eines Operators $o$ in einem Plan $\pi$.

**Definition 1** (Operator-Counting-Constraint)**.** *Sei $\Pi$ ein Planungsproblem mit Operatoren $\mathcal{O}$, und sei $s$ ein Zustand von $\Pi$. Sei $\mathcal{Y} = \{\mathsf{Count}_o \mid o \in \mathcal{O}\}$ eine Menge von nichtnegativen ganzzahligen Variablen.*[4] *Eine Menge von linearen Ungleichungen über $\mathcal{Y}$ ist ein* Operator-Counting-Constraint *für $s$, wenn für jeden $s$-Plan $\pi$ die Belegung $\mathsf{Count}_o = occur_o(\pi)$ alle Ungleichungen in der Menge erfüllt.*

Jeder Plan muss alle Operator-Counting-Constraints erfüllen, daher ist die billigste Möglichkeit, alle Constraints zu erfüllen, eine zulässige Heuristik.

**Definition 2** (Operator-Counting-Heuristik)**.** *Sei $\Pi$ ein Planungsproblem mit Operatoren $\mathcal{O}$ und der Kostenfunktion cost. Sei weiterhin $s$ ein Zustand von $\Pi$ und $C$ eine Menge von Operator-Counting-Constraints für $s$. Dann ist der Wert der* Operator-Counting-MIP-Heuristik $h^{\text{MIP}}$ *für $C$ und $s$ definiert durch das Integer-Programm:*

$$\min \left\{ \sum_{o \in \mathcal{O}} cost(o)\mathsf{Count}_o \;\middle|\; C \text{ ist erfüllt und } \mathsf{Count}_o \geq 0 \text{ für alle } o \in \mathcal{O} \right\}$$

*Wenn die Bedingung, dass die Variablen $\mathsf{Count}_o$ nur ganzzahlige Werte annehmen dürfen, aufgegeben wird, ergibt sich die LP-Relaxierung des Integer-Programms und wir sprechen von der* Operator-Counting-LP-Heuristik $h^{\text{LP}}$.

Die Berechnung von $h^{\text{LP}}$ ist polynomiell in der Größe von $C$, während die Berechnung von $h^{\text{MIP}}$ im Allgemeinen NP-äquivalent ist. Sowohl für LPs als auch für MIPs existieren in der Praxis hochoptimierte Programme. Beide Heuristiken sind zulässig, und durch Hinzufügen von Operator-Counting-Constraints können sich ihre Werte nur verbessern:

**Satz 2** (Heuristikeigenschaften [Po14])**.** *Für jede endliche Menge von Operator-Counting-Constraints $C$ gilt $h^{\text{LP}}_C \leq h^{\text{MIP}}_C \leq h^*$, und für jede endliche Menge von Operator-Counting-Constraints $C' \supseteq C$ gilt $h^{\text{LP}}_C \leq h^{\text{LP}}_{C'}$ und $h^{\text{MIP}}_C \leq h^{\text{MIP}}_{C'}$.*

Heuristiken, die sich als Operator-Counting-Heuristiken darstellen lassen, können somit leicht kombiniert werden, indem ihre Constraints vereinigt werden. Der Heuristikwert der Kombination ist mindestens so hoch ist wie das Maximum der beiden Heuristiken, und wir demonstrieren empirisch, dass er auch höher seien kann [Po14].

Wir zeigen außerdem, dass viele bekannte Heuristiken als Operator-Counting-Heuristiken darstellbar sind. Ein Beispiel sind Landmarken-Heuristiken [HD09]: Eine Landmarke ist

[4] Wir verwenden hier eine etwas vereinfachte Definition. Die vollständige Definition [Po14, Po17] erlaubt noch Hilfsvariablen, die für einige Heuristiken benötigt werden.

in diesem Zusammenhang eine Menge von Operatoren $L$, von denen mindestens einer verwendet werden muss. Dies lässt sich mit dem Constraint $\sum_{o \in L} \mathsf{Count}_o \geq 1$ leicht ausdrücken. Auch Heuristiken, die auf Delete-Relaxierung [IF15], Abstraktionsheuristiken [PRH13], Petri-Netzen [Bo13] und Netzwerkfluss [va07] beruhen, lassen sich auf diese Art darstellen [Po14, PHB17a]. Mit Ausnahme von Heuristiken, die auf kritischen Pfaden basieren [HG00], deckt dies alle bekannten Heuristikklassen ab. Für letztere beweisen wir, dass es unmöglich ist, sie als Operator-Counting-Heuristik auszudrücken [Po17].

## 6 Potentialheuristiken

Im dritten Teil der Dissertation stellen wir die Familie der *Potentialheuristiken* vor. So wie Operator-Counting-Heuristiken sind Potentialheuristiken *deklarativ*: Sie verwenden eine feste mathematische Form und erlauben es, gewünschte Eigenschaften der Heuristik als Constraints zu formulieren. Die Synthese der Heuristikfunktion wird dabei selbst zum Optimierungsproblem, das von einem spezialisierten Programm gelöst wird.

Potentialheuristiken verwenden numerische *Gewichte* für eine Menge von *Zustandseigenschaften* und berechnen das *Potential* eines Zustands als Summe der Gewichte aller Eigenschaften dieses Zustands. Ein Beispiel dafür ist der Materialwert einer Schachposition: Jede Zustandseigenschaft prüft, ob eine bestimmte Figur noch im Spiel ist, und hat ein entsprechendes Gewicht, z.B. 9 für die eigene Dame und $-3$ für einen gegnerischen Läufer.

Für Planungsprobleme verwenden wir Konjunktionen von Variablen-Wert-Paaren und sagen, dass ein Zustand $s$ die Eigenschaft $f = \bigwedge_{i=1}^{n} \langle V_i, v_i \rangle$ hat (geschrieben als $s \models f$), wenn $s(V_i) = v_i$ für alle $1 \leq i \leq n$. Die Zahl der Konjunktionsglieder $n$ ist die *Größe* von $f$. Wir verwenden die Iverson-Klammer $[P]$ für die Indikatorfunktion einer Eigenschaft $P$.

**Definition 3** (Potentialheuristik [Po15])**.** *Sei $\Pi$ ein Planungsproblem, $\mathscr{F}$ eine Menge von Zustandseigenschaften für $\Pi$ und $w : \mathscr{F} \to \mathbb{R}$ eine Gewichtsfunktion. Die* Potentialheuristik *für $\mathscr{F}$ und $w$ bildet jeden Zustand $s$ auf sein Potential ab:*

$$h^w(s) = \sum_{f \in \mathscr{F}} w(f)[s \models f]$$

*Die* Dimension *von $h^w$ ist die maximale Größe einer Eigenschaft $f \in \mathscr{F}$.*

Jede endliche Heuristik $h$ kann als Potentialheuristik mit ausreichend großer Dimension gesehen werden. Im Extremfall wird die Menge aller Zustände als Menge $\mathscr{F}$ verwendet, wobei jeder Zustand $s$ in eine Konjunktion der Größe $|\mathscr{V}|$ und dem Gewicht $h(s)$ ist. Wir sind daher besonders an Potentialheuristiken mit niedriger Dimension interessiert. Außerhalb der optimalen Handlungsplanung kann die Dimension von Potentialheuristiken verwendet werden, um die Komplexität von Problemen zu beschreiben [Se16]. Je kleiner die Dimension, die nötig ist, um eine gierigen Suche direkt zum Ziel zu führen, desto einfacher das Problem. Viele der üblichen Benchmarks benötigen dazu nur eine Dimension von 2, was bedeutet, dass solche Heuristiken auch für die optimale Planung interessant sind.

Zulässigkeit und Konsistenz sind zwei für die optimale Planung wichtige Heuristikeigenschaften und folgen aus den Ungleichungen $h^w(s_\star) \leq 0$ und $h^w(s) \leq cost(o) + h^w(s[\![o]\!])$ für

alle Zustände $s$ und darin anwendbare Operatoren $o$.[5] Die Anzahl dieser Ungleichungen ist exponentiell in der Größe des Planungsproblems. Für eindimensionale Zustandseigenschaften können sie aber kompakter aufgeschrieben werden:

$$\sum_{f \in \mathcal{F}} w(f)[s_\star \models f] \leq 0$$

$$\sum_{f \in pre(o) \cap \mathcal{F}} w(f) - \sum_{f \in eff(o) \cap \mathcal{F}} w(f) \leq cost(o) \quad \text{für alle } o \in \mathcal{O}$$

Diese Bedingungen sind notwendig und hinreichend dafür, dass eine eindimensionale Potentialheuristik zulässig und konsistent ist. Da außerdem alle Ungleichungen linear in den Gewichten sind, kann mit einem LP die *beste* solche Potentialheuristik in polynomieller Zeit gefunden werden, solange das Qualitätsmaß auch eine lineare Funktion ist. Zum Beispiel kann der durchschnittliche Heuristikwert oder $h(s_1)$ maximiert werden [SPH15].

Eine solche kompakte Charakterisierung existiert auch für die zulässigen und konsistenten zweidimensionalen Potentialheuristiken [PHB17b]. Für höherdimensionale Heuristiken ist dies aber unter der Annahme $\mathrm{P} \neq \mathrm{NP}$ unmöglich. Mit einer Erweiterung des Bucket-Elimination-Algorithmus [De03] können wir aber zeigen, dass der allgemeine Fall einfach (fixed parameter tractable) bleibt, solange die Abhängigkeit der Eigenschaften begrenzt ist.

In der Praxis liefern Potentialheuristiken gute Werte und sind schnell zu evaluieren, da die teure Optimierung der Gewichte nur einmal nötig ist. Sie lassen sich gut miteinander und mit anderen Heuristiken kombinieren und führen zu kompetitiven Ergebnissen [SPH15].

## 7 Zusammenhänge

Die drei Hauptteile der Dissertation sind enger miteinander verwandt, als es auf den ersten Blick erscheint. Sowohl Operator-Counting-Heuristiken als auch Potentialheuristiken können auf gewisse Weise als Kostenpartitionierung verstanden werden.

Mehrere Operator-Counting-Heuristiken können kombiniert werden, indem ihre Constraints vereinigt werden und dann in *einem* LP/MIP verwendet werden. Wir zeigen, dass die Kombination in einem LP genau der optimalen Kostenpartitionierung entspricht [Po15]:

**Satz 3.** *Seien $C_1, \ldots, C_n$ Mengen von Operator-Counting-Constraints für einen Zustand $s$, und sei $C = \bigcup_{i=1}^{n} C_i$. Dann ist $h^{\mathrm{LP}}_C(s)$ der Heuristikwert einer optimalen Kostenpartitionierung der Heuristiken $h^{\mathrm{LP}}_{C_1}, \ldots, h^{\mathrm{LP}}_{C_n}$, bei der negative Kosten erlaubt sind.*

Daraus folgt direkt, dass die Kombination mit $h^{\mathrm{LP}}_C$ besser ist als optimale nicht-negative Kostenpartitionierung und dass $h^{\mathrm{MIP}}_C$ sogar noch höhere Werte erreichen kann. Andersherum hilft Satz 3 auch dabei, Heuristiken besser zu verstehen: Operator-Counting-Heuristiken können nun als Kombination von kleineren Heuristiken interpretiert werden. So kann zum Beispiel die State-Equation-Heuristik [va07, Bo13] als Kombination von einfachen PDB-Heuristiken interpretiert und als solche noch verstärkt werden [PHB17a].

[5] Der Einfachheit halber nehmen wir hier an, dass das Ziel ein vollständiger Zustand ist, und dass jeder Operator die gleichen Variablen in Vorbedingung und Effekt erwähnt. Dies ist jedoch keine Einschränkung [PH15].

Zulässige und konsistente Potentialheuristiken können als Kostenpartitionierung von sogenannten *Flow-Heuristiken* betrachtet werden, die mit PDB-Heuristiken übereinstimmen, wenn unnötige Zustände aus den Abstraktionen entfernt werden [PHB17a].

**Satz 4.** *Sei $A = \langle \alpha_1, \ldots, \alpha_n \rangle$ eine Menge von Abstraktionsfunktionen, und sei $\mathscr{F}_A$ die Menge der abstrakten Zustände aus den von A induzierten Abstraktionen. Eine zulässige und konsistente Potentialheuristik für die Zustandseigenschaften $\mathscr{F}_A$ entspricht einer Transitionskostenpartitionierung der Flow-Heuristiken für $\alpha_1, \ldots, \alpha_n$, wobei negative Kosten erlaubt sind.*

Im Unterschied zu Operator-Counting-Heuristiken werden die Gewichte von Potentialheuristiken nur einmal berechnet und dann für alle Zustände verwendet. Für den Zusammenhang mit Kostenpartitionierung bedeutet das, dass eine Partitionierung gefunden wird, die dann für die ganze Suche verwendet wird, anstatt in jedem Zustand eine Partitionierung zu finden, die für diesen Zustand optimal ist. Die Partitionierung ist für das gleiche Qualitätsmaß optimal, für das die auch die Potentialheuristik optimiert wurde.

## Literaturverzeichnis


[BN95] Bäckström, Christer; Nebel, Bernhard: Complexity Results for $SAS^+$ Planning. Computational Intelligence, 11(4):625–655, 1995.

[Bo13] Bonet, Blai: An Admissible Heuristic for $SAS^+$ Planning Obtained from the State Equation. In: Proc. IJCAI 2013. S. 2268–2274, 2013.

[CS98] Culberson, Joseph C.; Schaeffer, Jonathan: Pattern Databases. Computational Intelligence, 14(3):318–334, 1998.

[De03] Dechter, Rina: Constraint Processing. Morgan Kaufmann, 2003.

[Ed01] Edelkamp, Stefan: Planning with Pattern Databases. In: Proc. ECP 2001. S. 84–90, 2001.

[GNT04] Ghallab, Malik; Nau, Dana; Traverso, Paolo: Automated Planning: Theory and Practice. Morgan Kaufmann, 2004.

[HBG05] Haslum, Patrik; Bonet, Blai; Geffner, Héctor: New Admissible Heuristics for Domain-Independent Planning. In: Proc. AAAI 2005. S. 1163–1168, 2005.

[HD09] Helmert, Malte; Domshlak, Carmel: Landmarks, Critical Paths and Abstractions: What's the Difference Anyway? In: Proc. ICAPS 2009. S. 162–169, 2009.

[HG00] Haslum, Patrik; Geffner, Héctor: Admissible Heuristics for Optimal Planning. In: Proc. AIPS 2000. S. 140–149, 2000.

[HHH08] Helmert, Malte; Haslum, Patrik; Hoffmann, Jörg: Explicit-State Abstraction: A New Method for Generating Heuristic Functions. In: Proc. AAAI 2008. S. 1547–1550, 2008.

[HNR68] Hart, Peter E.; Nilsson, Nils J.; Raphael, Bertram: A Formal Basis for the Heuristic Determination of Minimum Cost Paths. IEEE Transactions on Systems Science and Cybernetics, 4(2):100–107, 1968.

[IF15] Imai, Tatsuya; Fukunaga, Alex: On a Practical, Integer-Linear Programming Model for Delete-Free Tasks and its Use as a Heuristic for Cost-Optimal Planning. JAIR, 54:631–677, 2015.

[KD09] Karpas, Erez; Domshlak, Carmel: Cost-Optimal Planning with Landmarks. In: Proc. IJCAI 2009. S. 1728–1733, 2009.

[KD10] Katz, Michael; Domshlak, Carmel: Optimal admissible composition of abstraction heuristics. AIJ, 174(12–13):767–798, 2010.

[Ke16] Keller, Thomas; Pommerening, Florian; Seipp, Jendrik; Geißer, Florian; Mattmüller, Robert: State-dependent Cost Partitionings for Cartesian Abstractions in Classical Planning. In: Proc. IJCAI 2016. S. 3161–3169, 2016.

[Pe84] Pearl, Judea: Heuristics: Intelligent Search Strategies for Computer Problem Solving. Addison-Wesley, 1984.

[PH15] Pommerening, Florian; Helmert, Malte: A Normal Form for Classical Planning Tasks. In: Proc. ICAPS 2015. S. 188–192, 2015.

[PHB17a] Pommerening, Florian; Helmert, Malte; Bonet, Blai: Abstraction Heuristics, Cost Partitioning and Network Flows. In: Proc. ICAPS 2017. S. 228–232, 2017.

[PHB17b] Pommerening, Florian; Helmert, Malte; Bonet, Blai: Higher-Dimensional Potential Heuristics for Optimal Classical Planning. In: Proc. AAAI 2017. S. 3636–3643, 2017.

[Po14] Pommerening, Florian; Röger, Gabriele; Helmert, Malte; Bonet, Blai: LP-based Heuristics for Cost-optimal Planning. In: Proc. ICAPS 2014. S. 226–234, 2014.

[Po15] Pommerening, Florian; Helmert, Malte; Röger, Gabriele; Seipp, Jendrik: From Non-Negative to General Operator Cost Partitioning. In: Proc. AAAI 2015. S. 3335–3341, 2015.

[Po17] Pommerening, Florian: New Perspectives on Cost Partitioning for Optimal Classical Planning. Dissertation, Universität Basel, 2017.

[PRH13] Pommerening, Florian; Röger, Gabriele; Helmert, Malte: Getting the Most Out of Pattern Databases for Classical Planning. In: Proc. IJCAI 2013. S. 2357–2364, 2013.

[Se16] Seipp, Jendrik; Pommerening, Florian; Röger, Gabriele; Helmert, Malte: Correlation Complexity of Classical Planning Domains. In: Proc. IJCAI 2016. S. 3242–3250, 2016.

[SPH15] Seipp, Jendrik; Pommerening, Florian; Helmert, Malte: New Optimization Functions for Potential Heuristics. In: Proc. ICAPS 2015. S. 193–201, 2015.

[va07] van den Briel, Menkes; Benton, J.; Kambhampati, Subbarao; Vossen, Thomas: An LP-Based Heuristic for Optimal Planning. In: Proc. CP 2007. S. 651–665, 2007.

**Florian Pommerening** (*9.3.1985, Tübingen) studierte an der Universität Freiburg und schloss dort 2012 seinen Master in Informatik ab. Während dieser Zeit verbrachte er ein Jahr in Melbourne und beendete an der Monash University einen Master in Information Technology. Seit 2012 arbeitet er an der Universität Basel in der Gruppe von Prof. Dr. Malte Helmert im Bereich der klassischen Handlungsplanung. Nach seiner Promotion im Jahr 2017 führt er seine Arbeit als Postdoc in dieser Gruppe fort.

# Gestaltung und Analyse von räumlicher Navigation und geräteübergreifender Interaktion für das UbiComp[1]

Roman Rädle[2]

**Abstract:**

Diese Arbeit befasst sich mit der Gestaltung von räumlichen und geräteübergreifenden Interaktionstechniken. Als zentrale Themen präsentiert sie Forschung, die einerseits auf *Embodiment*-Praktiken basiert und andererseits, im Rahmen der Mensch-Computer-Interaktion bereits bestehende praktische Kenntnisse des täglichen Lebens ausnutzt. Diese *Embodiment*-Praktiken werden bei der täglichen Arbeit oft unbewusst angewandt und bieten neue—noch unerforschte—Potenziale für UbiComp Erfahrungen, die Spaß und Freude während der Bedienung bereiten sollen.

Derzeit erleben wir eine stark zunehmende Präsenz von leistungsstarken mobilen Geräten um uns herum. Geräte wie Smartphones und Tablets sind unsere alltäglichen Begleiter. Wenn wir sie nicht bereits in der Hand halten, dann warten sie oft in unseren Hosen-, Jacken- oder Tragetaschen, um uns überall und jederzeit mit ihrer Rechenleistung zu unterstützen (sog. Ubiquitous Computing oder kurz UbiComp [We91]).

Allerdings erkennen sie nur die Präsenz anderer Geräte, aber nicht deren genaue Lokation und sind daher sozusagen noch "blind". Somit ist auch das Ausführen von Aufgaben über Gerätegrenzen hinweg in der Regel umständlich, was dem Umstand geschuldet ist, dass Richtlinien für die Gestaltung von geräteübergreifenden Interaktionen (sog. cross-device interactions) fehlen.

Diese Arbeit schließt die oben erwähnte Lücke und befasst sich mit der Gestaltung von räumlichen und geräteübergreifenden Interaktionstechniken für das UbiComp. Als zentrale Themen präsentiert sie Forschung, die einerseits auf *Embodiment*-Praktiken basiert und andererseits im Rahmen der Mensch-Computer-Interaktion bereits bestehende praktische Kenntnisse des täglichen Lebens ausnutzt. Diese Embodiment-Praktiken werden bei der täglichen Arbeit oft unbewusst angewandt und bieten neue, noch unerforschte Potenziale für UbiComp Erfahrungen, die Spaß und Freude während der Bedienung bereiten sollen. Daher wurden im Rahmen dieser Arbeit zwei grundlegende Herausforderungen als Forschungsziele angegangen (sog. Research Objectives oder kurz RO, siehe Abb. 1):

- **RO1:** Benutzern soll eine Navigation und Interaktion in virtuellen Informationsräumen unter Verwendung bereits vorhandenen Wissens ermöglicht werden.
- **RO2:** Es sollen Möglichkeiten gesucht werden, die eine geräteübergreifende Interaktion nur unter Verwendung von handelsüblicher Hardware erlaubt.

[1] Englischer Titel der Dissertation: “Designing UbiComp Experiences for Spatial Navigation and Cross-Device Interactions” [Rä17a]

[2] Aarhus University, contact@romanraedle.com

Als Ausgangspunkt wird in Kapitel 2 das theoretische Fundament für diese Arbeit gelegt. Das Kapitel motiviert die grundlegende Vision einer Welt des “Ubiquitous Computing” und diskutiert verschiedene wissenschaftlich begründete Meinungen über den Erfolg und Misserfolg von UbiComp und nennt mögliche Gründe dafür. Gerade diese Diskussion zeigt die oft widersprüchlichen Ansichten von Forschern, insbesondere von Forschern der Mensch-Computer-Interaktion. Diese Ansichten werden unter Zuhilfenahme der sozialwissenschaftlichen Embodiment-Theorie [Do99], dem Reality-Based Interaction Framework [Ja08] und dem Blended Interaction Framework [JRG14] beleuchtet. Daraus ergeben sich erste Hinweise auf bisher ungenutzte Potentiale für UbiComp-Erfahrungen. Als Leitsatz gilt: der Benutzer steht im Vordergrund und nicht die Technologie. Der Benutzer kann bei Bedarf und jederzeit auf die Computertechnologie zugreifen, ohne sich von der Technologie diktieren zu lassen wie und wann sie verwendet werden soll [Ro06].

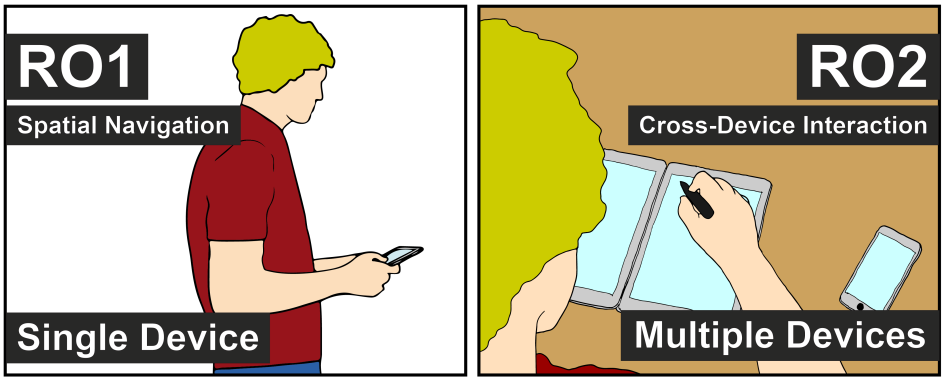



Abb. 1: Forschungsziele (RO) die in dieser Thesis bearbeitet werden: Räumliche Navigation (RO1) and geräteübergreifende Interaktion (RO2). Die Forschungsziele spannen von der Interaktion mit einem Gerät bis zur Interaktion mit mehreren Geräten.

Diese Potentiale für neue UbiComp-Erfahrungen werden in Kapitel 3 mittels Empirie weiter gestärkt. Hierin wird die Arbeit in den Kontext Wissensarbeit in wissenschaftlichen Bibliotheken gerückt und durchgeführte Feldstudien zur Datenerhebung vorgestellt. Die verschiedenen Analysen der Daten ergeben hilfreiche Erkenntnisse über die Nutzung von Computertechnologie während der Wissensarbeit und ermöglichen es Rückschlüsse darüber zu ziehen, welche Vorteile Anwender von computerbasierten Anwendungen erhalten, aber auch welchen Barrieren sie gegenüberstehen. Als zwei wichtige Erkenntnisse daraus ergeben sich die Notwendigkeit einer räumlichen Suche und Navigation in großen digitalen Bibliotheksbeständen und die freie Anordnung von digitalen Arbeitsmaterialien im Raum.

Die folgenden Kapitel 4-7 operationalisieren diese Potentiale und erforschen deren Nutzen in kontrollierten Experimenten. Sie versuchen außerdem, die Bedeutung des Raumes als kognitive Ressource zu verstehen, indem sie verschiedene räumliche und geräteübergreifende Interaktionen und deren Auswirkungen auf die Benutzerleistung (z. B. Navigation und Erinnerungsvermögen eines Benutzers) und subjektive Arbeitsbelastung (z. B. körper-

liche und geistige Beanspruchung) betrachten. Ihre Erkenntnisse werden in den folgenden Abschnitten präsentiert. Zunächst werden beide Forschungsziele, räumliche Navigation (RO1) und geräteübergreifende Interaktion (RO2), getrennt voneinander betrachtet und in einem integrativen Schritt zu einem Gesamtergebnis zusammengefasst.

# 1 Räumliche Navigation

Aufgrund ihrer großen Bildschirmfläche haben wandgroße Displays den Vorteil, dass Benutzer einen gesamten Informationsraum oder zumindest wesentliche Teile desselben auf einmal sehen können. Benutzer können zurücktreten, um den Inhalt der Anzeige zu überblicken und auf das Display zugehen, um Objekte von Interesse zu erkennen und darauf zuzugreifen. Sie sind ebenso nicht darauf angewiesen, Objektpositionen aus dem räumlichen Gedächtnis abzurufen. Große Displays unterstützen daher Erkennung statt Abruf (engl. recognition rather than recall[3]). Wie die Blended Shelf Benutzerstudie, in Kapitel 3, zeigt, führt ein großes Display aber auch zu einem Verlust der Privatsphäre. Inhärent unterstützen kleinere Bildschirme die Privatsphäre während der Erkundung und Navigation eines digitalen Informationsraums im Vergleich zu einem wandgroßen Display.

## 1.1 Peephole-Größe und Navigationsverhalten

Ein kleinerer Bildschirm hat auch seinen Preis. Benutzer müssen den Off-Screen-Inhalt unter Umständen mithilfe von Ansichtsverwaltungstechniken wie Multi-Touch-Navigation oder Peephole-Navigation manuell erkunden. Letztere ist eine zunehmend beliebte Technik zum Navigieren großer Informationsräume mittels kleinerer Bildschirme, die Inhalte abhängig von ihrer Position im Raum anzeigen [Fi93]. Dabei fungiert der Bildschirm als Fenster oder Guckloch (engl. Peephole) in einen viel größeren Informationsraum, z. B. eine Landkarte oder ein Bücherregal. Trotz dieses scheinbar offensichtlichen Nachteils der geringeren Bildschirmgröße zeigt eine Studie in Kapitel 4, dass ein relativ kleiner Peephole-Bildschirm in Tablet-Größe, im Vergleich zu einem wandgroßen Display, zu einer ähnlichen Aufgabenleistung für die Kartennavigation führt, wenn der Benutzer sich erst einmal mit dem Informationsraum vertraut gemacht hat. Dadurch verringert sich der Vorteil eines großen und oft kostspieligen Displays im Laufe der Zeit.

Die Größe des Bildschirms für Peephole Navigation kann verschiedene Ausprägungen annehmen und dementsprechend auch unterschiedliche Einflüsse auf die Leistung des Benutzers haben. Wie Kapitel 4 (Peephole-Größe und Navigationsverhalten) zeigt, verringert ein größerer Peephole-Bildschirm die Notwendigkeit für langsames physisches Schwenken (engl. panning) und Suchen und ermöglicht ein schnelleres visuelles Erfassen (engl. scanning) des Bildschirminhaltes. Es ermöglicht darüber hinaus eine Erkennung von Objekten im Informationsraum, anstatt des Abrufs derselben aus dem räumlichen Gedächt-

[3] "Minimize the user's memory load by making objects, actions, and options visible." (Nielsen Norman Group) — `https://www.nngroup.com/articles/ten-usabilityheuristics/` (zuletzt aufgerufen am 19. Januar 2018)

nis (engl. recognition rather than recall)[4]. In realen Systemen erhöhen größere Peephole-Bildschirme jedoch die Kosten, den Energieverbrauch und das Gewicht. Außerdem sind die Geräte mühsamer zu tragen und zu handhaben.

Designer müssen aufgrund dieser Einschränkungen Zugeständnisse machen. Sie möchten, dass Benutzer die Vorteile von größeren Peephole-Bildschirmen erleben und gleichzeitig die vielen Nachteile vermeiden, die sich aus der Verwendung und Handhabung größerer Geräte oder mobiler Projektionen ergeben. Die Beantwortung der Frage, "Wie klein dürfen Peephole-Bildschirme sein, ohne ihre Nutzer während der Navigation zu beeinträchtigen?", ist daher von hoher praktischer Relevanz für heutige mobile Augmented Reality (AR) Anwendungen (z. B. Pokémon Go). Es lässt sich außerdem in gewissem Maße auch auf die Nutzung von Mixed-Reality (MR) Hardware wie z. B. der Microsoft HoloLens übertragen. Hierbei besteht der Disput unter Herstellern und Anwendern gleichermaßen, wie gross muss das visuelle Sichtfeld (sog. field of view oder kurz FOV) für die Nutzer der MR Hardware sein?

Die Ergebnisse der Untersuchungen in dieser Arbeit zeigen, dass ein Peephole-Bildschirm in Tablet-Größe (in etwa 11 Zoll) bereits einen "sweet spot" zwischen der Größe des Bildschirms und der Navigationsleistung der Benutzer sowie deren Arbeitsbelastung darstellt (siehe Abb. 2). Ein Peephole-Bildschirm in Smartphone-Größe ist zu klein und wird von allen größeren Bildschirmen übertroffen. Es überrascht nicht, dass die Forschung in dieser Arbeit ergab, dass größere Peepholes die Lern- und die Navigationsgeschwindigkeit signifikant verbessern und die Arbeitsbelastung verringern. überraschend ist, dass der zusätzliche Nutzen eines größeren Bildschirms sich mit zunehmender Bildschirmgröße verringert und ein Peephole-Bildschirm größer als ein Tablet-Bildschirm sich in Bezug auf bessere Navigationsleistung oder geringere Arbeitsbelastung nicht mehr auszahlt. Dies steht im Widerspruch zu einigen Fitts' Law Peephole Target Acquisition Models (TAM). In den existierenden TAM wird das Verhalten des Benutzers mittels statischer Verfahren modelliert, jedoch werden subjektive Faktoren wie mentale Anforderung oder Frustration in den Modellen gänzlich ignoriert. In dem Verfahren in dieser Arbeit wird der Benutzer in die Gleichung mit einbezogen.

### 1.2 Auswirkung von Multi-Touch-Navigation und Peephole-Navigation auf Räumliches Gedächtnis

Eine Alternative zur Peephole-Navigation ist die Multi-Touch-Navigation, mit der Benutzer in virtuellen Informationsräumen navigieren können, indem sie herkömmliche Drag-to-Pan- und Pinch-to-Zoom-Touch-Gesten verwenden. Wie jedoch Kapitel 5 zeigt, verlieren Benutzer mit dieser Navigationstechnik häufig die globale Orientierung im großen Informationsraum[5]. Während sie bei der egozentrischen Peephole-Navigation von ihrer physischen Position zur virtuellen Position gelangen können. Diese räumliche Orientierung ermöglicht ihnen, ihre globale Orientierung im virtuellen Informationsraum beizube-

[4] Teile dieses Kapitels wurden in wissenschaftlichen Beiträgen publiziert: [Rä14c]
[5] Teile dieses Kapitels wurden in wissenschaftlichen Beiträgen publiziert: [Rä13]

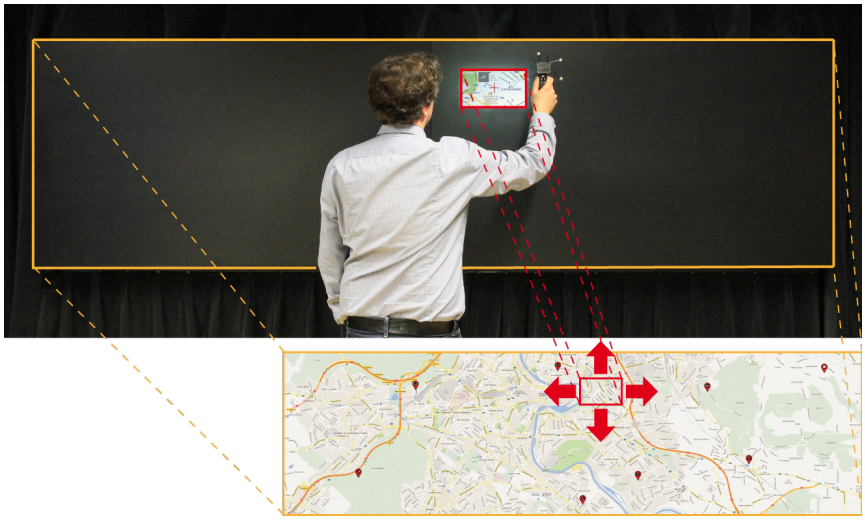


Abb. 2: Experimenteller Versuchsaufbau zur Simulation einer Dynamic Peephole Interaction auf einem großen vertikaler Bildschirm. Diese Simulation erlaubt es, die Wirkung von Peehhole-Navigation auf Kartennavigation mit hoher interner Validität. Es vermeidet Störfaktoren wie Gewicht und Auflösung bestimmter Geräte.

halten. Außerdem navigieren sie mit einer egozentrischen Peephole-Navigation effizienter im Informationsraum als mit Multi-Touch-Navigation.

Folglich ist die Peephole-Navigation eine echte Alternative zu wandgroßen Bildschirmen, vor allem dann wenn Privatsphäre ein Muss ist. Sie ist auch der traditionellen Multi-Touch-Navigation überlegen, führt zu einer besseren Navigationsleistung und reduziert die kognitiven Anforderungen an die Benutzer. Es geht sogar so weit, dass die Peephole-Interaktion das langfristige räumliche Gedächtnis, im Vergleich zur traditionellen Multi-Touch-Interaktion, besser ausnutzt.

In einer weiteren Studie, die in Kapitel 5 beschrieben ist, wird die Dynamic Peephole Navigation einer traditionellen Multi-touch Navigation gegenübergestellt. Hierbei wird auf den Vorarbeiten aus Kapitel 4, dem “sweet-spot” Tablet-Bildschirm, aufgebaut. Als Aufgabe mussten die Teilnehmer in einer großen virtuellen Karte navigieren und mehrfach dieselben Orte auf der Karte aufrufen. Es zeigt sich eine signifikant bessere Navigationsleistung für die Peephole-Navigation in Bezug auf die Pfadlänge (47 %) und die Aufgabenbearbeitungszeit (34 %). Darüber hinaus berichten die Teilnehmer von einer besseren Benutzererfahrung, einer deutlich geringeren mentalen Anforderung und Frustration während der Bewältigung der Aufgabe. So benutzten sie z. B. ihre physische Position oft als räumliche Orientierung, um ihre globale Orientierung im virtuellen Informationsraum aufrechtzuerhalten.

Daher empfiehlt sich die Verwendung von Peephole-Interaktion mit Tablets, um in großen virtuellen Informationsräumen zu navigieren, wenn Datenschutz, Navigationsleistung und

kognitive Anforderungen wichtige Voraussetzungen sind. Dies ist insbesondere in solchen Kontexten von Bedeutung, in denen der Benutzer erhebliche kognitive Ressourcen für Aufgaben- auf Anwendungsebene investieren muss, z. B. während zeitkritischen Entscheidungen, Wissenskonstruktion oder allgemeinen Lernaufgaben höherer Ordnung. Dies muss jedoch mit einer höheren physischen Belastung durch die körperliche Bewegung in Einklang gebracht werden, wenn Navigationsvorgänge sehr häufig und über einen längeren Zeitraum ausgeführt werden und somit zu körperlicher Belastung und Ermüdung führen.

## 2 Geräteübergreifende Interaktion

Die Notwendigkeit einer parallelen und sequentiellen Nutzung mehrerer Mobilgeräte wurde von Forschern [JOO15, Ce16] und Industrie[6] gleichermaßen aufgedeckt. Die Integrative Workplace Benutzerstudie in Kapitel 3 (Kontext und Analyse) bestätigt diese Notwendigkeit. Wissensarbeiter ordnen ihre Arbeitsartefakte räumlich auf ihren Arbeitsbereichen an, um Dokumente zu vergleichen oder Querverweise zu erstellen. Weiterhin stellt die Interaktion über mehrere persönliche Geräte eines Nutzers hinweg oder das übertragen von Information von einem Gerät zu einem Gerät einer anderen Person ein gut dokumentiertes Problem in der Literatur dar [SW13, JOO15]. Trotz der berichteten täglichen sequenziellen und parallelen Nutzung mobiler Geräte bietet die Mensch-Computer-Interaktion noch immer wenig Unterstützung für geräteübergreifende Interaktionen. Schlimmer noch, die Fragen nach dem geeigneten Design von geräteübergreifenden Interaktionen und ob diese räumlich bewusster sein müssen, bleiben bisher unbeantwortet. Es mangelt an Leitprinzipien für geräteübergreifende Interaktionen [Ou08].

Zudem erlauben bisherige Technologien für geräteübergreifende Interaktion entweder keine räumliche Interaktion [HW14] oder erfordern eine aufwendige Instrumentierung von mobilen Geräten [Sc10] oder Räumen mit teuren Tracking-Systemen [MHG12]. Darüber hinaus sind die meisten geräteübergreifenden Interaktionssysteme Forschungsprototypen und Closed-Source- Systeme und somit für die Durchführung von ökologisch validen Nutzerstudien nicht verfügbar.

### 2.1 Tracking Technologie für geräteübergreifende Interaktion

Kapitel 6 beschreibt HuddleLamp[7], eine Sensortechnologie in Form einer Schreibtischlampe mit integrierter RGB-D-Kamera, die die Bewegungen mehrerer mobiler Displays auf einem Tisch verfolgt. HuddleLamp ermöglicht verschiedene Arten von Szenarien. Zum Beispiel Einzelbenutzerszenarien wie die Interaktion über mehrere Geräte hinweg um gleichzeitiges Lesen und Schreiben von Dokumenten zu ermöglichen. Es ermöglicht auch die Zusammenarbeit zwischen mehreren Benutzern und ihren mobilen Geräten (siehe Abb. 3).

[6] The new multi-screen world study: `https://www.thinkwithgoogle.com/advertising-channels/mobile/the-new-multi-screen-world-study/` (zuletzt aufgerufen am 19. Januar 2018)

[7] Teile dieses Kapitels wurden in wissenschaftlichen Beiträgen publiziert: [Rä14b, Rä14a]

HuddleLamp[8] ist das erste Tracking-System für mobile Geräte, das mit einem vertretbaren Aufwand an Umgebungsinstrumentierung auskommt und mit handelsüblicher Hardware arbeitet. Hierzu reicht eine kostengünstige RGB-D Kamera, die über einem Tisch, z. B. in einer Schreibtischlampe, montiert wird. Das Tracking-System erkennt mehrere mobile Displays durch Ausnutzung ihrer optischen Eigenschaften im IR-Bereich. Es verwendet zusätzlich RGB Bilder in einem neuartigen Hybrid-Sensing Verfahren, um die Tracking Zuverlässigkeit zu steigern. Mithilfe des Hybrid-Sensing kann es Bildschirme von Hintergrund, anderen Objekten oder den Händen der Benutzer unterscheiden und über die Zeit hinweg erkennen.

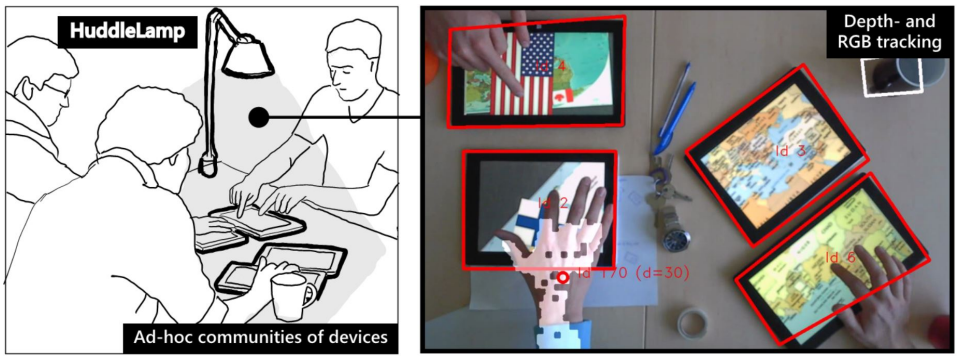



Abb. 3: Das HuddleLamp Tracking-System erkennt und verfolgt mobile Geräte und die Hände der Benutzer und ermöglicht dadurch eine Ad-hoc-Zusammenarbeit mehrerer Benutzer.

Die webbasierte Architektur von HuddleLamp Anwendungen ermöglicht den Nutzern überdies mobile Geräte ad hoc hinzufügen oder wieder zu entfernen ohne zusätzliche Software installieren zu müssen. In fünf Beispielen wird der praktische Einsatz von HuddleLamp demonstriert, u. a. der Einsatz um große Multi-Device-Displays für Multi-User- und Multi-Touch-Interaktionen zu erstellen. über die fünf vorgestellten Beispiele hinaus ermöglicht HuddleLamp die Erforschung zukünftiger geräteübergreifender Interaktionen.

## 2.2 Räumliche geräteübergreifende Interaktion verstehen

Eine Studie in Kapitel 7 exploriert in einem zweistufigen Verfahren den “Designspace” von mobilen geräteübergreifenden Interaktionen[10]. Insbesondere um Antworten und Lösungen auf Probleme zu finden, die in Kapitel 3 (Kontext und Analyse) aufgeworfen wurden. In der ersten Stufe stellt es Ergebnisse aus einer Gesture-Elicitation-Study vor. Hierzu wurden 17 Teilnehmer zu geräteübergreifender Interaktionen befragt. Sie sollten Vorschläge zu, in der Literatur, dokumentierten geräteübergreifenden Interaktionen geben. 71 % der Vorschläge waren räumlich. Dies deutet darauf hin, dass die Teilnehmer der Studie bei geräteübergreifenden Interaktionen vorzugsweise räumlich denken (vgl. [Ki10]). In

[8] Die HuddleLamp Software ist online verfügbar als Open-Source-Projekt und kann für Forschungszwecke frei eingesetzt werden: [9] (zuletzt aufgerufen am 19. Januar 2018).

[10] Teile dieses Kapitels wurden in wissenschaftlichen Beiträgen publiziert: [Rä15]

der zweiten Stufe und basierend auf den Vorschlägen der Studienteilnehmer wurden zwei räumliche geräteübergreifende Interaktionstechniken (Edge Bubbles und Radar View) und eine räumlich-agnostische Interaktionstechnik (Menu) implementiert und diese in einem kontrollierten Experiment verglichen. Die Ergebnisse zeigen, dass räumliche geräteübergreifende Interaktionstechniken von den Benutzern bevorzugt werden und sich ihre mentale Anforderung, Anstrengung und Frustration dabei verringern kann. Es ist jedoch wichtig zu erwähnen, dass das "Design" einer räumlichen geräteübergreifenden Interaktionstechnik eine signifikante Rolle bei (i) Benutzerpräferenz und (ii) der subjektiven Arbeitsbelastung der Benutzer spielt.

Abschließend werden in dieser Arbeit die einzelnen Erkenntnisse zusammengefasst und in allgemeingültige Gestaltungsrichtlinien für zukünftige räumliche und geräteübergreifende Anwendungen überführt. Diese Richtlinien können von Forschern und Praktikern angewendet werden, um neue UbiComp Interaktionsformen und Nutzererfahrungen zu entwickeln. Diese Interaktionsformen sollen letztlich die Frustration des Nutzers senken und gleichzeitig seine Leistungsfähigkeit steigern.

Fortführende Arbeiten [Pl17, Rä17b, Rä18, Pa18].

## Literaturverzeichnis

[Ce16] Cecchinato, Marta E; Sellen, Abigail; Shokouhi, Milad; Smyth, Gavin: Finding Email in a Multi-Account, Multi-Device World. In: Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems - CHI '16. 2016.

[Do99] Dourish, Paul: Embodied Interaction: Exploring the Foundations of a New Approach to HCI. 1999.

[Fi93] Fitzmaurice, George W.: Situated information spaces and spatially aware palmtop computers. Communications of the ACM, 36(7):39–49, jul 1993.

[HW14] Hamilton, Peter; Wigdor, Daniel J.: Conductor: enabling and understanding cross-device interaction. In: Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14. CHI '14, ACM Press, New York, New York, USA, S. 2773–2782, apr 2014.

[Ja08] Jacob, Robert J.K.; Girouard, Audrey; Hirshfield, Leanne M.; Horn, Michael S.; Shaer, Orit; Solovey, Erin Treacy; Zigelbaum, Jamie: Reality-based interaction: a framework for post-WIMP interfaces. In: Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08. ACM Press, New York, New York, USA, S. 201, 2008.

[JOO15] Jokela, Tero; Ojala, Jarno; Olsson, Thomas: A Diary Study on Combining Multiple Information Devices in Everyday Activities and Tasks. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15. ACM Press, New York, New York, USA, S. 3903–3912, apr 2015.

[JRG14] Jetter, Hans-Christian; Reiterer, Harald; Geyer, Florian: Blended Interaction: understanding natural human-computer interaction in post-WIMP interactive spaces. Personal and Ubiquitous Computing, 18(5):1139–1158, jun 2014.

[Ki10] Kirsh, David: Thinking with external representations. AI & Society, 25(4):441–454, feb 2010.

[MHG12] Marquardt, Nicolai; Hinckley, Ken; Greenberg, Saul: Cross-device interaction via micro-mobility and f-formations. In: Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12. ACM Press, New York, New York, USA, S. 13, oct 2012.

[Ou08] Oulasvirta, Antti: When users "do"the Ubicomp. interactions, 15(2):6, mar 2008.

[Pa18] Park, Seonwook; Gebhardt, Christoph; Rädle, Roman; Feit, Anna Maria; Vrzakova, Hana; Dayama, Niraj Ramesh; Yeo, Hui-Shyong; Klokmose, Clemens N.; Quigley, Aaron; Oulasvirta, Antti; Hilliges, Otmar: AdaM: Adapting Multi-User Interfaces for Collaborative Environments in Real-Time. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, ACM, New York, NY, USA, S. 184:1–184:14, 2018.

[Pl17] Plank, Thomas; Jetter, Hans-Christian; Rädle, Roman; Klokmose, Clemens N.; Luger, Thomas; Reiterer, Harald: Is Two Enough?: Studying Benefits, Barriers, and Biases of Multi-Tablet Use for Collaborative Visualization. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. CHI '17, ACM, New York, NY, USA, S. 4548–4560, 2017.

[Rä13] Rädle, Roman; Jetter, Hans-Christian; Butscher, Simon; Reiterer, Harald: The effect of egocentric body movements on users' navigation performance and spatial memory in zoomable user interfaces. In: Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces - ITS '13. ACM Press, New York, New York, USA, S. 23–32, 2013.

[Rä14a] Rädle, Roman; Jetter, Hans-Christian; Marquardt, Nicolai; Reiterer, Harald; Rogers, Yvonne: Demonstrating HuddleLamp: Spatially- Aware Mobile Displays for Ad-hoc Around-the-Table Collaboration. In: Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces - ITS '14. ACM Press, New York, New York, USA, S. 435–438, nov 2014.

[Rä14b] Rädle, Roman; Jetter, Hans-Christian; Marquardt, Nicolai; Reiterer, Harald; Rogers, Yvonne: HuddleLamp: Spatially-Aware Mobile Displays for Ad-hoc Around-the-Table Collaboration. In: Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces - ITS '14. ACM Press, New York, New York, USA, S. 45–54, 2014.

[Rä14c] Rädle, Roman; Jetter, Hans-Christian; Müller, Jens; Reiterer, Harald: Bigger is not always better: display size, performance, and task load during peephole map navigation. In: Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14. ACM Press, New York, New York, USA, S. 4127–4136, apr 2014.

[Rä15] Rädle, Roman; Jetter, Hans-Christian; Schreiner, Mario; Lu, Zhihao; Reiterer, Harald; Rogers, Yvonne: Spatially-aware or Spatially-agnostic?: Elicitation and Evaluation of User-Defined Cross-Device Interactions. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15. ACM Press, New York, New York, USA, S. 3913–3922, 2015.

[Rä17a] Rädle, Roman: Designing UbiComp Experiences for Spatial Navigation and Cross-Device Interactions. Dissertation, University of Konstanz, 2017.

[Rä17b] Rädle, Roman; Nouwens, Midas; Antonsen, Kristian; Eagan, James R.; Klokmose, Clemens N.: Codestrates: Literate Computing with Webstrates. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. UIST '17, ACM, New York, NY, USA, S. 715–725, 2017.

[Rä18] Rädle, Roman; Jetter, Hans-Christian; Fischer, Jonathan; Gabriel, Inti; Klokmose, Clemens N.; Reiterer, Harald; Holz, Christian: PolarTrack: Optical Outside-In Device Tracking That Exploits Display Polarization. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, ACM, New York, NY, USA, S. 497:1–497:9, 2018.

[Ro06] Rogers, Yvonne: Moving on from Weiser's Vision of Calm Computing: engaging UbiComp experiences. 4206:404–421, sep 2006.

[Sc10] Schmitz, Arne; Li, Ming; Schönefeld, Volker; Kobbelt, Leif: Ad-Hoc Multi-Displays for Mobile Interactive Applications. In: The Eurographics Association. The Eurographics Association, S. 45–52, 2010.

[SW13] Santosa, Stephanie; Wigdor, Daniel: A field study of multi-device workflows in distributed workspaces. In: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp '13. ACM Press, New York, New York, USA, S. 63, 2013.

[We91] Weiser, Mark: The computer for the 21st century. Scientific American, 265(3):94–104, sep 1991.

**Roman Rädle** ist Assistant Professor in der Abteilung für Computer Science an der Universität Aarhus in Dänemark. Zuvor war er Postdoctoral Fellow in der Abteilung für Digitales Design und Informationswissenschaften ebenfalls an der Universität Aarhus. Er hat einen B.Sc., M.Sc. und Dr. rer. nat. in Informatik von der Universität Konstanz in Deutschland verliehen bekommen.

Er veröffentlicht regelmäßig auf renomierten akademischen Konferenzen wie der CHI, UIST und ISS. Er ist Mitglied des SIGCHI Operations Committee, das logistische, technische und operative Fragen behandelt, die SIGCHI und dessen Portfolio von Konferenzen betreffen.

Im Jahre 2013/2014 war er Gastwissenschaftler am Game Innovation Lab an der New York University. Von August bis Dezember 2015 war er während eines 4,5-monatigen Praktikums bei Microsoft Research in Cambridge in die Gruppe Human Experience & Design eingebettet. Seine Forschungsinteressen umfassen Mensch-Computer-Interaktion, Ubiquitous Computing und Computational Thinking. Er arbeitet an der Verwendung interaktiver Notizbücher in Situationen, in denen die Anzahl der Personen und Geräte im Laufe der Zeit variieren kann. Sowohl Software also auch Hardware sollen die übergänge zwischen diesen Situationen fließend und ohne mentale Gymnastik unterstützen. Zum Beispiel ihre Verwendung an Schulen, in denen Arbeiten mit Notizbüchern von Einzelarbeit ("Ich-Arbeit") zu kollaborativer Gruppenarbeit ("Wir-Arbeit") und umgekehrt wechseln kann.

# Algorithmen für datenintensive Graph- und Clusteringprobleme[1]

Chris Schwiegelshohn[2]

**Abstract:** Thema der vorliegenden Dissertation sind algorithmische Aspekte der modernen Datenanalyse, wobei die einzelne Analyse eines Datensatzes auf der Optimierung einer Zielfunktion basiert. Trotz einer langen Forschungshistorie in diesem Bereich gibt es hier immer noch zahlreiche offene Probleme. Außerdem traten um die Jahrtausendwende neue Fragestellungen auf, da Datensätze inzwischen oft so groß sind, dass sich viele der bisherigen Optimierungsverfahren nicht mehr anwenden lassen. Die Beiträge der Dissertation betreffen beide Aspekte.

Die folgenden Seiten fassen die Ergebnisse der Dissertation kurz zusammen. Diese werden sowohl miteinander in Beziehung gesetzt als auch in die bereits veröffentlichte Literatur eingeordnet. Das Augenmerk liegt dabei weniger auf den technischen Einzelheiten als auf der Frage, was sich ein interessierter Laie unter den Ergebnissen vorstellen sollte.

## Einleitung

Da der Begriff "Big Data" über das Fachgebiet Informatik hinaus Einzug in den täglichen Sprachgebrauch gefunden hat, ist es nicht verwunderlich, dass er auf vielerlei oft unterschiedliche Arten verwendet wird. Ein gemeinsames Charakteristikum lässt sich aber hinter allen seinen Nutzformen erkennen: Big Data befasst sich mit Datenmengen, die auf Grund ihres Volumens sowohl die Speicherung als auch die Analyse der Daten vor neue Herausforderungen stellen. Obwohl die Datenanalyse als Fachrichtung schon sehr lange existiert – bereits Carl Friedrich Gauß beschäftigte sich mit linearer Regression – rückten viele wichtige Fragestellungen in diesem Bereich erst im Zusammenhang mit Big Data in den Fokus der Forschung, insbesondere auch in der theoretischen Informatik.

Bis zu einem gewissen Grad lässt sich eine Vergrößerung von Datenmengen durch die Verwendung von mehreren Speicher- und Recheneinheiten zusammen mit Parallelverarbeitung handhaben. Bei sehr großen Datenmengen reicht der Einsatz mehrerer physischer Einheiten jedoch nicht mehr aus. Dann werden Algorithmen benötigt, die die Datensätze verkleinern, wobei der entstehende komprimierte Datensatz den Ausgangsdatensatz gut repräsentieren muss. Außerdem werden vielfach so genannte *Datenstromalgorithmen* eingesetzt, die den Datensatz sequentiell Datum um Datum verarbeiten und dabei die Kompression berechnen. Wenn die geplante Verwendung des Datensatzes bereits vor seiner Generierung bekannt ist, kann die Kompression mit dem klassischen Problem der Datenanalyse, der Extraktion von Informationen aus dem Datensatz, kombiniert werden. Besonders interessant sind dabei Problemstellungen, die sich in der Praxis für kleinere Datenmengen bereits bewährt haben. Dazu gehören insbesondere Clusteringprobleme.

[1] Englischer Titel der Dissertation: "On Algorithms for Large-Scale Graph and Clustering Problems"

[2] Sapienza University of Rome, schwiegelshohn@diag.uniroma1.it

## Ergebnisse der Dissertation

Die Dissertation enthält im Wesentlichen fünf Ergebnisse, die einzeln in den folgenden Veröffentlichungen dargestellt wurden:

- Marc Bury and Chris Schwiegelshohn. Sublinear Estimation of Weighted Matchings in Dynamic Data Streams. European Symposium of Algorithms (ESA), 2015 [BS15]
- Marc Bury, Chris Schwiegelshohn, and Mara Sorella. Sketch 'em all: Approximate Similarity Search on Dynamic Data Streams. Conference on Web Search and Data Mining (WSDM), 2018 [BSS18]
- Vincent Cohen-Addad, Chris Schwiegelshohn, and Christian Sohler. Diameter and k-Center in Sliding Windows. International Colloquium on Automata, Languages, and Programming, (ICALP) 2016 [CSS16]
- Marc Bury and Chris Schwiegelshohn. On Finding the Jaccard Center. International Colloquium on Automata, Languages, and Programming, (ICALP) 2017 [BS17]
- Vincent Cohen-Addad and Chris Schwiegelshohn. On the Local Structure of Stable Clustering Instances. Symposium on Foundations of Computer Science, (FOCS) 2017 [CS17]

Bei den ersten beiden Veröffentlichungen handelt es sich um Datenstromalgorithmen, die letzten beiden Veröffentlichungen beschreiben Lösungsverfahren von Clusteringproblemen und die dritte Veröffentlichung schlägt die Brücke zwischen diesen beiden Gebieten.

## Grundlagen zur Clusteranalyse in großen Graphen

Die Dissertation basiert auf der Grundannahme, dass der Datensatz in Form eines Graphen mit Elementarobjekten, den Knoten, und ihren binären Verknüpfungen, den Kanten, beschrieben wird. Eine Kante kann eine Vielzahl von Bedeutungen haben. In sozialen Netzwerken, in denen Nutzer durch Knoten repräsentiert werden, bedeutet eine Kante, dass zwei Nutzer befreundet sind. Handelt es sich bei den Knoten dagegen um Punkte etwa im Euklidischen Raum, gibt die gewichtete Kante die Distanz zwischen den beiden Punkten an. Clusteringprobleme gibt es in zu vielen Variationen, um sie an dieser Stelle einzeln aufzuzählen. Das grundlegende Konzept lässt sich aber wie folgt einfach darstellen:

> Beim Clustering soll eine gegebene Menge von Objekten so in Gruppen, auch "Cluster" genannt, eingeteilt werden, dass
>
> - ähnliche Objekte dem gleichen Cluster und
> - unähnliche Objekte verschiedenen Cluster zugeordnet werden.

Gemäß dieser Charakterisierung wird also ein Ähnlichkeits- oder Distanzmaß (im Folgenden mit sim oder dist abgekürzt) benötigt, um ein Clusteringproblem zu beschreiben.

Dabei wird das Distanzmaß häufig so gewählt, dass es eine Metrik beschreibt, siehe zum Beispiel die Euklidische Distanz, die für zwei gegebene Punkte $x = (x_1, x_2)$ und $y = (y_1, y_2)$ die Form $\text{dist}(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$ annimmt.

Für ungewichtete Graphen wird eine andere, aber dennoch verwandte Metrik benutzt. Dabei werden zwei Knoten als ähnlich bezeichnet, wenn sie eine große gemeinsame Nachbarschaft haben. Dies wird unter anderem durch den Jaccard-Koeffizient $\text{sim}(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$ modelliert, wobei für einen Knoten $u$ die Menge $N(u)$ alle mit $u$ verknüpfte Knoten beinhaltet. Mit Hilfe der Beziehung $\text{dist}(u,v) = 1 - \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$ ergibt sich die Jaccard-Distanz aus dem Jaccard-Koeffizienten. Tatsächlich lässt sich die Jaccard Distanz in Euklidische Räume einbetten [GL86], was bedeutet, dass Jaccard Distanzen lediglich eine spezielle, eingeschränkte Auswahl von Euklidischen Distanzen sind.

Mittels solcher Ähnlichkeits- und Distanzmaße erfolgt die Bewertung eines Clusterings. Auch hier gibt es eine Vielzahl an Möglichkeiten. Einer der beliebtesten Ansätze ist das zentrumsbasierte Clustering, bei dem jedes Cluster neben der ihm zugeordneten Objekte ein Zentrumsobjekt beinhaltet. Die Güte eines Clusterings ist dann eine Funktion der Distanzen von den Objekten $P$ der Cluster $C$ zu den jeweiligen Zentren $c$. Häufig verwendete Funktionen sind

**($k$-Median)** $\sum_{\text{Cluster } C_i} \sum_{x \in C_i} \text{dist}(x, c_i)$,

**($k$-Means)** $\sum_{\text{Cluster } C_i} \sum_{x \in C_i} \text{dist}^2(x, c_i)$ und

**($k$-Center)** $\max_{\text{Cluster } C_i} \max_{x \in C_i} \text{dist}(x, c_i)$.

Wird die Anzahl der Cluster auf $k$ beschränkt, beschreiben diese drei Zielfunktionen das $k$-median, das $k$-means und das $k$-center Problem. Der zweifellos bekannteste Repräsentant dieser Probleme ist das $k$-means Problem in Euklidischen Räumen. Dies wird besonders dann deutlich, wenn man sich die algebraischen Eigenschaften des Problems vor Augen führt: Sei $P$ eine Menge von Punkten im Euklidischen Raum, $c$ ein beliebiger Punkt und $\mu$ der Zentroid von $P$ definiert als $\frac{1}{|P|} \sum_{x \in P} x$. Dann gilt folgende Gleichung:

$$\sum_{x \in P} \text{dist}^2(x,c) = \sum_{x \in P} \text{dist}^2(x,\mu) + |P| \cdot \text{dist}^2(\mu,c) \tag{1}$$

Wird $P$ als eine Matrix mit $n$ Zeilen und $d$ Spalten betrachtet, so entsprechen die Zeilen den Koordinaten eines Punktes. Mit der Frobeniusnorm $\|P\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{d} P_{i,j}^2}$ der Matrix $P$ lässt sich Gleichung (1) wie folgt umformulieren:

$$\|P - C\|_F^2 = \left\| P - \left(\frac{\mathbf{1}}{\sqrt{\mathbf{n}}}\right)\left(\frac{\mathbf{1}}{\sqrt{\mathbf{n}}}\right)^T P \right\|_F^2 + \left\| C - \left(\frac{\mathbf{1}}{\sqrt{\mathbf{n}}}\right)\left(\frac{\mathbf{1}}{\sqrt{\mathbf{n}}}\right)^T P \right\|_F^2,$$

wobei $C$ die Matrix ist, deren Zeilen nur aus dem Punkt $c$ bestehen, und $\left(\frac{\mathbf{1}}{\sqrt{\mathbf{n}}}\right)$ der $n$-dimensionale Vektor ist, bei dem alle Einträge den Wert $\frac{1}{\sqrt{n}}$ haben. Es sei darauf hingewiesen, dass $\left(\frac{\mathbf{1}}{\sqrt{\mathbf{n}}}\right)\left(\frac{\mathbf{1}}{\sqrt{\mathbf{n}}}\right)^T P$ die Zeilen von $P$ auf den Zentroiden $\mu$ von $P$ abbildet. Dies lässt sich auf mehrere Cluster durch die $n \times k$ Clustermatrix $X$ verallgemeinern:

$$X_{i,j} = \begin{cases} \frac{1}{\sqrt{|C_j|}} & \text{wenn } P_i \in C_j \\ 0 & \text{sonst} \end{cases}.$$

Damit ergibt sich folgende Formulierung des $k$-means Problems:

$$\min_{\text{Clustermatrix } X} \|P - XX^T P\|_F^2 \qquad \text{oder} \qquad \max_{\text{Clustermatrix } X} \|XX^T P\|_F^2.$$

Die algebraischen Eigenschaften von $k$-means sind eng verwandt mit denen eines weiteren Problems aus der Dissertation, dem *Maximum Matching*. Ausgangspunkt ist ein ungewichteter Graph. Ein Matching $M$ ist eine Auswahl an Kanten, so dass keine zwei verschiedenen Kanten aus $M$ einen gemeinsamen Knoten haben. Insbesondere für die Spieltheorie ist Maximum Matching ein zentrales Problem, es findet allerdings auch Anwendung für $k$-median, $k$-means und $k$-center Probleme [Ch17]. Obwohl Maximum Matching rein kombinatorisch definiert und bearbeitet werden kann, hat es ebenfalls eine starke algebraische Komponente. Algebraische Ansätze zum Maximum Matching basieren auf der sogenannten Tutte-Matrix

$$T_{i,j} = \begin{cases} x_{i,j} & \text{wenn } i > j \text{ und Kante } (i,j) \text{ existiert} \\ -x_{i,j} & \text{wenn } j > i \text{ und Kante } (i,j) \text{ existiert} \\ 0 & \text{sonst,} \end{cases},$$

wobei die $x_{i,j}$ Variablen sind. Tutte konnte zeigen, dass ein Matching für alle Knoten möglich ist, wenn es eine Wahl der Variablen von $T$ gibt, so dass die Determinante von $T$ nicht 0 ist. Lovász [Lo79] hat dies noch verallgemeinert, indem er bewies, dass der maximal mögliche Rang von $T$ über alle Variablenbelegungen genau der doppelten Größe des Maximum Matching entspricht. Ferner zeigte er, dass eine zufällige Belegung der Variablen mit hoher Wahrscheinlichkeit zu einer Matrix mit maximalem Rang führt.

Sowohl Rang als auch Frobeniusnorm sind Funktionen über dem Spektrum der Matrix. In gewisser Weise lassen sich somit die Approximation des $k$-means Problems und die Approximation der Matchinggröße als Spezialfälle von Matrixapproximationen auffassen. Obwohl diese Sichtweise nicht unbedingt die am weitesten verbreitete Herangehensweise an das Matchingproblem ist, gewinnen algebraische Methoden für Matching zunehmend an Bedeutung. Tatsächlich ist das Resultat in der Dissertation einer der ersten Datenstromalgorithmen, der durch $k$-means inspirierte algebraische Methoden auf das Matchingproblem anwendet.

## Approximation der Matchinggröße

Maximum Matching ist ein immer noch sehr intensiv untersuchtes Problem auf Graphen, siehe McGregor [Mc14] für einen Überblick, und zweifellos das Graphenproblem, das im Datenstrom die größte Aufmerksamkeit erhalten hat. Da ein Matching in einem Graphen bestehend aus $n$ Knoten eine maximale Größe von bis zu $n/2$ haben kann, benötigen alle Algorithmen mindestens $\Omega(n)$ Speicher um ein großes Matching zu berechnen. Datenstromalgorithmen für das Maximum Matching streben daher an $O(n \text{ polylog } n)$ Speicher zu verwenden. Es ist allerdings kein Algorithmus bekannt, der ohne annähernd alle Kanten abzuspeichern, weniger als eine 2-Approximation garantieren kann und $O(n \text{ polylog } n)$ Platz ermöglicht bestenfalls eine $\frac{e}{e-1}$-Approximation [Ka13]. In dynamischen Datenströmen, die sowohl das Hinzufügen als auch das Löschen von Kanten erlauben, müssen sogar im Wesentlichen alle Kanten abgespeichert werden.

Trotz ihrer häufigen Verwendung hat die $O(n \text{ polylog } n)$ Platzschranke den Nachteil, dass große Graphen aus der Praxis, wie planare Graphen und Grad beschränkte Graphen, weniger Kanten haben. Für solche Graphen sind Algorithmen mit $O(n \text{ polylog } n)$ Platzbedarf unbrauchbar. Da das Matching jedoch $\Omega(n \log n)$ Platz benötigt, entstand Interesse an der Approximation der Größe des Matchings im Datenstrom. Meist wurde sich dabei auf das normale Datenstrommodell beschränkt, bei dem Kanten nur hinzugefügt werden.

In der Dissertation wurde erstmals gezeigt, dass sich eine sublineare $o(n)$ Platzschranke auch auf dynamische Datenströme, bei denen Kanten gelöscht werden, übertragen lässt. Die zentrale Idee basiert auf Rang-erhaltenden Einbettungen, die auf die Tutte-Matrix angewendet werden. Solche Rang-erhaltenden Einbettungen wurden schon für das $k$-means Problem benutzt [Co15]. Als Einbettungen werden Rademacher-Matrizen verwendet, deren Einträge uniform zufällig auf $-1$ oder 1 gesetzt werden.

Ferner wurde in der Dissertation bewiesen, dass sich jeder Datenstromalgorithmus zur Approximation der Matchinggröße zur Lösung eines schwierigen Problems aus der Kommunikationskomplexität benutzen lässt, woraus sich untere Schranken für das Matching-Problem ergeben.

## Clustering und Datenstromalgorithmen für Jaccard-Koeffizienten

Der Jaccard-Koeffizient spielt eine große Rolle in der Web-getriebenen Datenanalyse. Das Jaccard-Center Problem besteht darin, für eine gegebene Kollektion $N$ von Mengen über einer Grundgesamtheit $U$ eine Teilmenge $C \subset U$ zu finden, so dass $\max_{X \in N} 1 - \frac{|X \cap C|}{|X \cup C|}$ minimiert wird. Das Problem ist eine Variante des kleinsten-umschließenden-Ball-Problems für Euklidische Räume. Motiviert wird es durch die Anwendung des Jaccard-Koeffizienten in der Plagiatserkennung [Br97]: Ein plagiatisierter Text und das Original haben einen hohen Jaccard-Koeffizienten. Hierbei wird für eine gegebene Menge an Texten die Ähnlichkeit durch den Jaccard-Koeffizienten bezüglich der vorkommenden Worte und Phrasen ermittelt. Die bedeutsamsten Phrasen ihrerseits können durch eine Lösung des 1-Center Problems bestimmt werden.

Neben der Lösung des Optimierungsproblems ist die Auswertung des Jaccard-Koeffizienten einer der zeitlichen Hauptengpässe. Um schnell viele Texte auf Plagiate zu überprüfen, werden spezielle Hashfunktionen verwendet. Anstatt alle aufgeführten Phrasen einzeln für jeden Text zu betrachten, werden sie in zufälliger Weise angeordnet und für jeden Text wird die jeweils erste vorkommende Phrase abgespeichert. Die Wahrscheinlichkeit, dass zwei Texte die gleiche Phrase besitzen, entspricht dann dem Jaccard-Koeffizienten. Das Verfahren ist in der Literatur unter dem Begriff Minhashing bekannt.

In der Dissertation werden einige Dimensionsreduktionsverfahren entwickelt, die für beide Arten von Problemen angewendet werden können. Für das 1-Center Optimierungsproblem ermöglichen diese Verfahren eine $(1+\varepsilon)$-Approximation in polynomieller Zeit für jedes feste $\varepsilon > 0$. Für das Auswertungsproblem wurde erstmals eine Form von Minhashing in dynamischen Datenströmen durchgeführt.

## Durchmesser und $k$-Center

Temporale Datenströme, auch als Datenströme mit Zeitfenster bezeichnet, bilden eine weitere Form der Veränderung von Datensätzen. Während bei dynamischen Datenströmen Daten gelöscht werden, existieren bei temporalen Datenströmen keine explizite Löschoperationen, sondern es werden immer nur die $n$ zuletzt eingefügten Daten betrachtet.

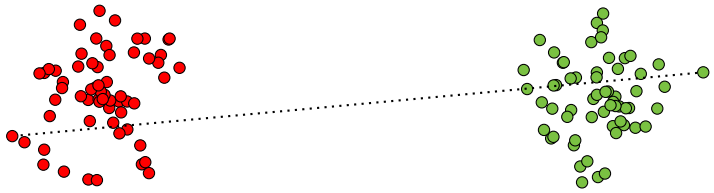


Abb. 1: Gegeben ist hier eine Punktmenge im zweidimensionalen Euklidischen Raum. Die beiden Punkte, die den Durchmesser induzieren, sind durch die gestrichelte Linie miteinander verbunden. Die Verwendung dieser beiden Punkte als Zentren für das 2-Center Problem garantiert eine 2-Approximation.

Für das 2-Center Problem wurde ein Algorithmus mit einem Approximationsfaktor 4 entdeckt. Dieses Ergebnis ist optimal, da ebenfalls gezeigt werden konnte, dass $\Omega(\sqrt{n})$ Punkte notwendig sind um die Schranke von 4 zu durchbrechen. Da ferner für normale Datenströme ein Approximationsfaktor von 2 optimal ist, ist dieses Ergebnis eine der ersten Separierungen von temporalen und normalen Datenströmen.

Der Algorithmus für das 2-Center Problem lässt sich ebenfalls für das Durchmesserproblem anwenden, siehe Abbildung 1 für eine grafische Darstellung der Beziehung beider Probleme. Beim Durchmesserproblem ist für eine gegebene Menge an Punkten der größten Abstand zwischen zwei Punkten finden. Die kritische Unterroutine verifiziert für eine gegebene Schätzung $\gamma$ des Durchmesser, ob es zwei Punkte mit dem Abstand mindestens $\gamma$ gibt. Falls der Algorithmus keine zwei solche Punkte finden kann, lässt sich zeigen, dass

der Durchmesser höchsten $3\gamma$ ist. Ein besonderer Augenmerk gilt der Platzkomplexität dieses Algorithmus, der mit nur 4 Punkten auskommt.

Für das allgemeine $k$-Center Problem lässt sich mit ähnlichen Techniken eine 6-Approximation zeigen. Ob dies verbessert werden kann oder eine abermalige Separierung von 2- auf 3-Center vorliegt, ist zu diesem Zeitpunkt noch offen.

## Lokale Suche und warum Clustering einfach ist

Eine beliebig gute Approximation ist sowohl für $k$-median als auch für $k$-means im Allgemeinen NP-schwer zu berechnen. Aus Sicht der theoretischen Informatik zählen beide Probleme daher zu den schwierigsten Problemen der Datenanalyse. Diese Einschätzung deckt sich allerdings nicht mit Erfahrungen aus der Praxis: Dort werden Clusteringverfahren zur Problemlösung erfolgreich eingesetzt, unter anderem auch als Unterprogramme für die praktisch effiziente Bearbeitung von Problemen, für die sogar Algorithmen mit polynomieller Laufzeit existieren. Diese scheinbare Diskrepanz hat ihre Ursache in der Ableitung des in der Theorie benutzte Schwierigkeitsbegriff aus dem schlimmstmöglichen Fall (Englisch: worst case) eines Problems. Oft wird daher die Brauchbarkeit des schlimmstmöglichen Falles für die Praxis sowohl allgemein als auch speziell im Bereich Clustering in Zweifel gezogen.

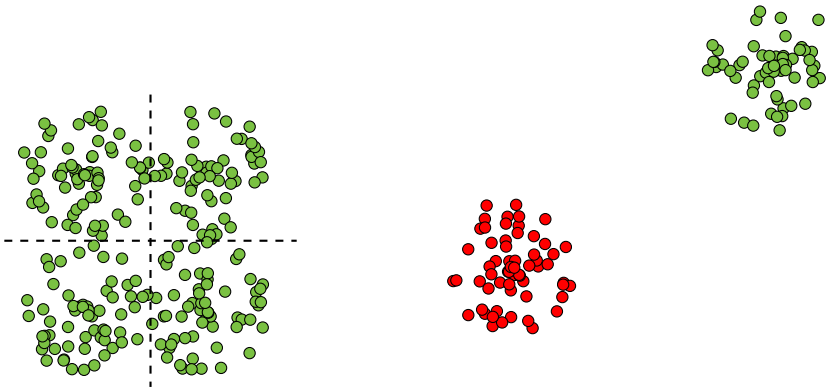


Abb. 2: Beispiele für ein schweres und ein offensichtliches 2-Clustering

Als Beispiel werden folgende zwei Instanzen für $k$-means Clustering mit $k = 2$ betrachtet, siehe Abbildung 2. In der linken Instanz gibt es kein offensichtliches 2-Clustering. Für die durch die horizontal oder durch vertikal gestrichelte Linie jeweils induzierte Partition gelten folgende zwei Beobachtungen:

- In beiden Fällen hat das Clustering nahezu identische Kosten.

- In beiden Fällen hat das Clustering eine Übereinstimmung von 50%.

Da eine Übereinstimmung von weniger als 50% bei zwei Clustern nicht möglich ist, sind die Cluster also maximal unähnlich bezüglich der Punktzugehörigkeit und verhalten sich gleichzeitig nahezu identisch bezüglich ihrer $k$-means Kosten. Im Gegensatz dazu ist das "korrekte"Clustering der rechten Instanz offensichtlich (und farbig gekennzeichnet).

Dieses Beispiel verdeutlicht die unterschiedlichen Herangehensweisen an Clusteringprobleme: In der Theorie steht die Optimierung einer Zielfunktion im Vordergrund. In der Praxis wird eine Zielfunktion so gewählt, dass sie (a) eine Grundannahme des Datensatzes modelliert und (b) leicht interpretierbar ist, wenn diese Grundannahme zutrifft. Auf das Beispiel bezogen bedeutet dies, dass die 2-means Zielfunktion nur für die rechte nicht aber für die linke Instanz brauchbar ist.

Für den Algorithmenentwurf führten diese und ähnliche Überlegungen zu einem Umdenken. Falls ein Algorithmus beispielsweise interpretierbare Ergebnisse für die rechte Instanz liefert, ist ein schlechtes Verhalten für die linke Instanz akzeptabel oder sogar unerheblich. Um solche Algorithmen zu entwickeln, wurden Einschränkungen für die zu betrachtenden Instanzen bestimmt, die uninteressante schlimmstmögliche Fälle vernachlässigen. Die Formulierung einer solchen Einschränkung ist allerdings nicht offensichtlich. Daher gibt es mehrere konkurrierende Modelle, siehe Abbildung 3.

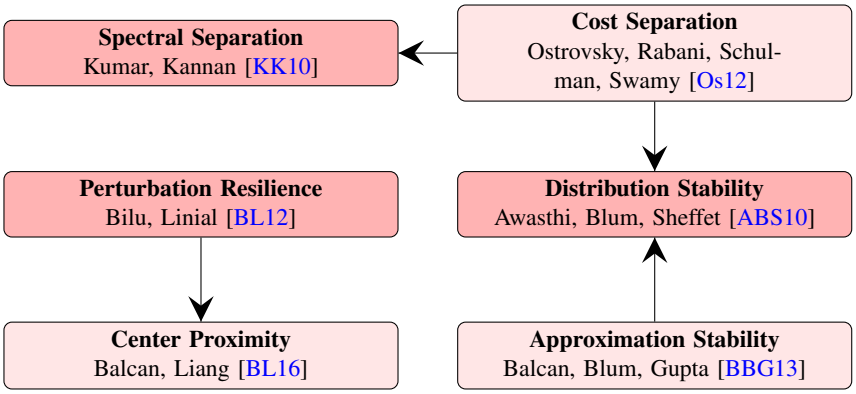



Abb. 3: Diese Grafik gibt einen Auswahl an Kriterien, die gut clusterbare Instanzen abbilden. Pfeile entsprechen einer Implikation: Zum Beispiel ist jede kostenseparierte Instanz (Cost Separation) ebenfalls spektral separiert. Die drei dunkel eingefärbten Definitionen haben in der Literatur die meiste Beachtung gefunden.

Die Theorie hinter diesen Separationseigenschaften hat einige Erfolge vorzuweisen. Beispielsweise wurden viele der in der Praxis eingesetzten Heuristiken für diverse Eigenschaften analysiert mit dem Ergebnis, dass ihr Abschneiden viel besser ist als durch die Analyse des schlimmstmöglichen Falles vorhergesagt wurde. Es sind jedoch noch viele Fragen offen: Zum einen sind viele der Separationseigenschaften untereinander unvergleichbar, während es im Idealfall nur ein Kriterium gibt, welches für die Praxis relevant

ist. Zum anderen versagen die bisher analysierten Heuristiken für mindestens eine der Separationseigenschaften.

Der Beitrag der Dissertation zu diesem Thema hat als Fokus die lokale Suchheuristik. Die lokale Suchheuristik ist streng genommen kein Clusteringverfahren sondern ein generischer Algorithmus, der auf viele Probleme angewendet wird [AL97]. Beim Clustering geht das Verfahren von einer gegebenen Menge an Zentren $C$ aus. Falls es eine Menge $C'$ gibt, die sich in ihrer Zusammensetzung nur geringfügig von $C$ unterscheidet und die Kosten des Clusterings verbessert, wechselt die lokale Suche auf $C'$. Dies wird so lange wiederholt, bis keine Veränderung mehr eintritt. Für einen Pseudocode sei auf Algorithmus 1 verwiesen.

**Algorithmus 1** Lokale Suche

1: $C \leftarrow$ beliebige Auswahl von $k$ Punkten
2: **while** $(\exists C'$ so dass $(|C' \triangle C| < p) \bigwedge (\text{Kosten}(C') < \text{Kosten}(C)))$ **do**
3: $\quad C \leftarrow C'$
4: **end while**

In der Dissertation wurde gezeigt, dass die lokale Suche für *alle* benutzten Stabilitätskriterien ebenfalls das Optimum oder eine $(1+\varepsilon)$-Approximation des Optimums ermittelt. Außerdem erzeugt die lokale Suche bereits gute Ergebnisse für Clustering im schlimmstmöglichen Fall, was für die Analyse von großer Bedeutung ist. Zusammenfassend lassen sich die Ergebnisse zur lokalen Suche wie folgt interpretieren.

- Eine Instanz ist gut clusterbar, wenn ihre lokalen Optima nah bei dem globalen Optimum liegen.
- Die Analyse des schlimmstmöglichen Falles ist wertvoll, selbst wenn diese Fälle nicht von dem Algorithmus berücksichtigt werden müssen.

## Literaturverzeichnis

[ABS10] Awasthi, Pranjal; Blum, Avrim; Sheffet, Or: Stability Yields a PTAS for k-Median and k-Means Clustering. In: FOCS. 2010.

[AL97] Aarts, Emile; Lenstra, Jan K., Hrsg. Local Search in Combinatorial Optimization. John Wiley & Sons, Inc., New York, NY, USA, 1st. Auflage, 1997.

[BBG13] Balcan, Maria-Florina; Blum, Avrim; Gupta, Anupam: Clustering under approximation stability. J. ACM, 60(2):8, 2013.

[BL12] Bilu, Yonatan; Linial, Nathan: Are Stable Instances Easy? Combinatorics, Probability & Computing, 21(5):643–660, 2012.

[BL16] Balcan, Maria-Florina; Liang, Yingyu: Clustering under Perturbation Resilience. SIAM J. Comput., 45(1):102–155, 2016.

[Br97] Broder, Andrei Z.; Glassman, Steven C.; Manasse, Mark S.; Zweig, Geoffrey: Syntactic Clustering of the Web. Computer Networks, 29(8-13):1157–1166, 1997.

[BS15] Bury, Marc; Schwiegelshohn, Chris: Sublinear Estimation of Weighted Matchings in Dynamic Data Streams. In: ESA. 2015.

[BS17] Bury, Marc; Schwiegelshohn, Chris: On Finding the Jaccard Center. In: ICALP. 2017.

[BSS18] Bury, Marc; Schwiegelshohn, Chris; Sorella, Mara: Sketch'Em All: Fast Approximate Similarity Search in Dynamic Data Streams. In: WSDM. 2018.

[Ch17] Chierichetti, Flavio; Kumar, Ravi; Lattanzi, Silvio; Vassilvitskii, Sergei: Fair Clustering Through Fairlets. In: NIPS. 2017.

[Co15] Cohen, Michael B.; Elder, Sam; Musco, Cameron; Musco, Christopher; Persu, Madalina: Dimensionality Reduction for k-Means Clustering and Low Rank Approximation. In: STOC. 2015.

[CS17] Cohen-Addad, Vincent; Schwiegelshohn, Chris: On the Local Structure of Stable Clustering Instances. In: FOCS. 2017.

[CSS16] Cohen-Addad, Vincent; Schwiegelshohn, Chris; Sohler, Christian: Diameter and k-Center in Sliding Windows. In: ICALP. 2016.

[GL86] Gower, J. C.; Legendre, P.: Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification, 3(1):5–48, 1986.

[Ka13] Kapralov, Michael: Better bounds for matchings in the streaming model. In: SODA. 2013.

[KK10] Kumar, Amit; Kannan, Ravindran: Clustering with Spectral Norm and the k-Means Algorithm. In: FOCS. 2010.

[Lo79] Lovász, László: On determinants, matchings, and random algorithms. In: FCT. S. 565–574, 1979.

[Mc14] McGregor, Andrew: Graph stream algorithms: a survey. SIGMOD Record, 43(1):9–20, 2014.

[Os12] Ostrovsky, R.; Rabani, Y.; Schulman, L. J.; Swamy, C.: The effectiveness of Lloyd-type methods for the k-means problem. J. ACM, 59(6):28, 2012.

**Chris Schwiegelshohn** wurde in München geboren und ist in Peekskill, New York, Dortmund und Herdecke zur Schule gegangen. Nach dem Abitur hat er an der Technischen Universität Dortmund Informatik studiert und anschließend promoviert. Inzwischen arbeitet er als Post-Doc an der römischen Universität Sapienza. Sein Hauptarbeitsgebiet ist die theoretische Informatik mit einem Fokus auf algorithmische Aspekte von großen Datenmengen. Daneben arbeitet er auch an anwendungsorientierten Problemen mit Bezug zu Maschinellem Lernen, Datamining und Webanwendungen.

# Argumentative Schreibunterstützung durch maschinelle Sprachverarbeitung[1]

Christian Stab[2]

**Abstract:** Die automatische Analyse von geschriebenen Argumenten ist wegen der hohen Ambiguität und Vagheit von natürlicher Sprache eine große Herausforderung. Zum einen erfordert das Trainieren von maschinellen Lernverfahren hochqualitative Korpora, und zum anderen besteht die automatische Erkennung von Argumentationsstrukturen aus mehreren voneinander abhängigen Analyseschritten. In dieser Dissertation wird die Anwendbarkeit von theoretischen Argumentationsmodellen auf argumentative Aufsätze erstmals untersucht und ein Korpus zur ganzheitlichen Erkennung von Argumentationsstrukturen erstellt. Es wird ein neues Modell zur Erkennung von Argumentationsstrukturen vorgestellt, das die Funktion von Argumentkomponenten und argumentativen Relationen global optimiert und die Erkennungsraten gegenüber lokalen Modellen verbessert. Zusätzlich wird ein Ansatz zur automatischen Qualitätsbewertung vorgestellt, mit dem unzureichend begründete Argumente in argumentativen Aufsätzen mit hoher Präzision erkennt werden.

## 1 Einleitung

Das Schreiben von argumentativen Aufsätzen ist eine effektive Methode, Argumentationsfähigkeiten zu lehren. Durch die Analyse kontroverser Meinungen lernen Schüler_innen, Argumente zur Begründung ihres eigenen Standpunkts zu formulieren, Gegenargumente zu berücksichtigen und logische Fehler zu vermeiden. Wegen des enormen Korrekturaufwands können Lehrkräfte jedoch nur eine geringe Anzahl an argumentativen Schreibaufgaben vergeben, die in der Praxis nicht ausreichen, um Schüler_innen ausreichend auszubilden [BB11]. Computergestützte Assistenz- oder Bewertungssysteme [SB13] liefern zwar Rückmeldungen zu Grammatik, Diskursstrukturen, und lexikalischem Umfang, sind aber noch nicht in der Lage natürlichsprachliche Argumente automatisch zu erkennen und zu bewerten. „Argument Mining“ – ein junges Forschungsfeld der automatischen Sprachverarbeitung – hat das Potential diese Lücke zu schließen und neue intelligente Schreibhilfen zu ermöglichen, die automatisch konstruktive Rückmeldungen zu natürlichsprachlichen Argumenten generieren.

Die automatische Erkennung und Bewertung von natürlichsprachlichen Argumenten unterliegt jedoch den folgenden Herausforderungen: (1) Methoden zur automatischen Analyse von Argumenten basieren – wie viele andere Verfahren der automatischen Sprachverarbeitung – auf überwachten maschinellen Lernmethoden. Diese lernen die Erkennung und Bewertung von Argumenten anhand von bekannten Argumenten in manuell annotierten Korpora. Wegen der Ambiguität und Vagheit von Texten ist eine eindeutige Interpretation

[1] Englischer Titel der Dissertation: „Argumentative Writing Support by means of Natural Language Processing“

[2] Technische Universität Darmstadt, Ubiquitous Knowledge Processing Lab (UKP-TUDA), Department of Computer Science, Hochschulstraße 10, 64289 Darmstadt, <nachname>@ukp.informatik.tu-darmstadt.de

von natürlichsprachlichen Argumenten selbst für Menschen eine schwierige Aufgabe. Zu Beginn dieser Arbeit war es weitestgehend unbekannt, ob existierende theoretische Argumentationsmodelle von menschlichen Annotatoren mit ausreichender Übereinstimmung auf argumentative Aufsätze angewandt werden können, um qualitativ hochwertige Trainingsdaten zu erstellen. (2) Die meisten existierenden Methoden zur automatischen Argumentextraktion adressieren spezifische Teilaufgaben wie die Identifikation von Argumentkomponenten, die Klassifikation der argumentativen Funktion von Argumentkomponenten (bspw. als Behauptung oder Prämisse), oder die Erkennung von argumentativen Relationen. Diese Aufgaben sind jedoch nicht unabhängig voneinander. Beispielsweise lässt sich die Funktion einer Argumentkomponente erst bestimmen, wenn die argumentative Diskursstruktur bekannt ist und umgekehrt. Daher ist die unabhängige Modellierung dieser Teilaufgaben nicht ausreichend, um konsistente Argumentationsstrukturen zu erkennen. (3) Die Bewertung von Argumenten ist eine höchst subjektive Aufgabe. Die Qualität eines Arguments ist das Produkt einer Vielzahl von Kriterien, die nicht nur von persönlichen Präferenzen sondern auch vom Vorwissen eines Individuums abhängig sind [Th73]. Zum Beispiel unterliegt die Qualität eines Arguments dem Grad des Vertrauens in den Argumentierenden (Ethos), den durch das Argument angesprochenen Emotionen (Pathos), der logischen Korrektheit der Argumentation (Logos) und der Situation in der das Argument geäußert wurde (Kairos) [SN13]. Die hohe Subjektivität stellt eine wesentliche Herausforderung für die Entwicklung von automatischen Methoden zur Qualitätsbewertung von Argumenten dar. Basierend auf diesen Herausforderungen adressiert diese Dissertation die folgenden drei Forschungsfragen (engl. „research questions“):

**RQ1** **Annotation von Argumentationsstrukturen**: Es sollte untersucht werden, ob sich theoretische Argumentationsmodelle auf studentische Aufsätze anwenden lassen. In diesem Zusammenhang wollen wir evaluieren, ob und in welchem Ausmaß menschliche Annotatoren bei der Erkennung von Argumentationsstrukturen übereinstimmen und ob es möglich ist, Trainingsdaten von hoher Qualität zu erstellen.

**RQ2** **Automatische Erkennung von Argumentationsstrukturen**: Die zweite Forschungsfrage adressierte die automatische Erkennung von Argumentationsstrukturen. Es sollte untersucht werden, welche linguistischen Eigenschaften für die Teilschritte der Argumenterkennung effektiv sind und ob die gemeinsame Modellierung der Analyseschritte die Genauigkeit verbessert.

**RQ3** **Automatische Qualitätsbewertung von Argumenten**: Mit dieser Forschungsfrage sollte beantwortet werden, welche Qualitätskriterien verwendet werden können, um objektives Feedback zu generieren, ob diese Kriterien auf echte Argumente anwendbar sind, und mit welcher Genauigkeit sie automatisch bewertet werden können.

Mit der Beantwortung dieser Forschungsfragen trägt diese Dissertation zu einem besseren Verständnis von natürlichsprachlichen Argumenten bei. Zum einen ermöglicht die Untersuchung der Anwendbarkeit von theoretischen (meist normativen) Argumentationsmodellen auf studentische Aufsätze eine empirische Beurteilung der Machbarkeit von automatischen Methoden zur Argumentanalyse. Andererseits leistet diese Dissertation methodische Beiträge zur automatischen Erkennung und Bewertung von Argumenten in Texten,

welche die Entwicklung von neuen Basistechnologien für innovative Suchverfahren zur Entscheidungsunterstützung ermöglichen.

## 2 Annotation von Argumentationsstrukturen

Die erste Herausforderung bestand in der Definition eines geeigneten Annotationsschemas. Dazu wurden in einem erste Schritt theoretische Ansätze aus Philosophie und der (informellen) Logik untersucht.[3] Um einen möglichst hohen Detailgrad der automatischen Analyse zu ermöglichen, lag der Schwerpunkt hierbei auf *monologischen Argumentationsmodellen*, die die Mikrostruktur und Komponenten von Argumenten formalisieren [BMB10]. Nach diesen Modellen besteht ein Argument aus mehreren Komponenten [Go10]: Die *Behauptung* ist eine kontroverse Aussage und die zentrale Komponente eines Arguments. Die *Prämissen* sind Aussagen zur Begründung oder Widerlegung der Behauptung. Die *Konsequenzrelation* verbindet die Prämisse(n) mit der Behauptung und legt die Art der Schlussfolgerung fest.

Der Vergleich existierender Argumentationsmodelle zeigte, dass *Argumentdiagramme* wegen ihrer Ähnlichkeit zu existierenden Diskursanalyseansätzen am Besten geeignet sind. Argumentdiagramme sind eine Methode aus der informellen Logik mit der unstrukturierte Argumente in natürlicher Sprache für die weitere Analyse in eine strukturierte Repräsentation überführt werden [Go10]. Ein Argumentdiagram besteht aus Knoten und gerichteten Kanten. Jeder Knoten entspricht einer Komponente des Arguments (eine Aussage in natürlicher Sprache). Eine Kante stellt eine gerichtete argumentative Beziehung von der Ursprungskomponente zur Zielkomponente dar (bspw. „begründet“ oder „widerlegt“). Im Gegensatz zu anderen normativen Argumentmodellen, wie beispielsweise dem Toulmin Modell [To58] oder den Argumentationsschemata von Walton [WRM08], können mit Argumentdiagrammen auch komplexe Argumentationsketten und geschachtelte Widerlegungen modelliert werden. Durch die explizite Modellierung von argumentativen Relationen können zudem mehrere Argumente in einem Text leicht voneinander getrennt werden.

### 2.1 Annotationsschema

Das vorgeschlagene Annotationsschema modelliert die Argumentstruktur eines Textes als Baum, wobei der Wurzelknoten die argumentative *Kernaussage* (engl. „major claim“) enthält. Die eigentlichen Argumente werden mit Knoten vom Typ `Behauptung` und `Prämisse` modelliert. Gerichtete argumentative Relationen vom Type `begründet` und `widerlegt` modellieren die Zusammenhänge zwischen den Argumentkomponenten. Der folgende Auszug eines Aufsatzes zum Thema Auslandssemester illustriert das vorgeschlagene Annotationsschema (Kernaussagen sind fett gedruckt, Behauptungen unterstrichen, und Prämissen gewellt unterstrichen):

> **Beispiel 1)**: „*Studieren im Ausland ist ein häufig diskutiertes Thema. Ich denke,* ***[jeder sollte während des Studiums ein Semester im Ausland verbringen]***$_{Kernaussage}$. *[Ein Auslandssemester*

[3] vgl. Kapitel 2 in [St17].

*ist eine wertvolle Erfahrung]$_{Behauptung}$. [Das Leben an einem fremden Ort wird natürlich erstmal schwierig sein]$_{Prämisse1}$, aber [diese Schwierigkeiten werden später zur wertvollen Erfahrung]$_{Prämisse2}$. [Man lernt außerdem auf eigenen Füßen zu stehen und nicht von anderen abhängig zu sein]$_{Prämisse3}$.“*

Der erste Satz leitet das Thema ein und enthält keinen argumentativen Inhalt. Der zweite Satz enthält die Kernaussage und die Meinung des Autors zum Thema Auslandssemester. Der dritte Satz ist die Behauptung des ersten Arguments zur Untermauerung der Kernaussage. Der vierte Satz enthält zwei Prämissen von denen $\texttt{Prämisse}_1$ die Behauptung und $\texttt{Prämisse}_2$ $\texttt{Prämisse}_1$ widerlegt. Der letzte Satz enthält eine weitere Prämisse welche die Begründung direkt belegt. Abbildung 1 zeigt die Argumentationsstruktur des Beispiels.

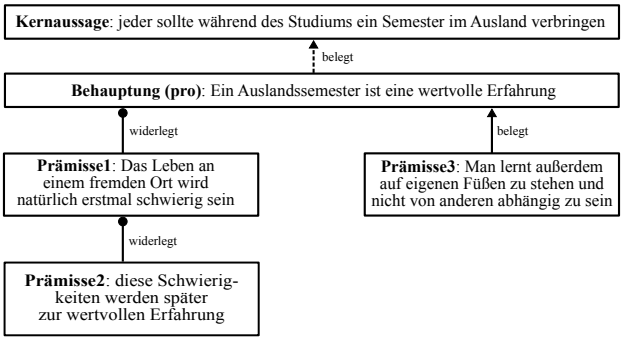



Abb. 1: Argumentationsstruktur von Beispiel 1.

## 2.2 Annotationsstudie und Erstellung der Trainingsdaten

Um die Anwendbarkeit des Annotationsschemas zu evaluieren wurde eine Annotationsstudie durchgeführt. Das Korpus für diese Studie ist eine Sammlung von 402 englischen Aufsätzen, die einem Onlineforum entnommen wurden.[4] Die Aufsätze wurden zufällig aus dem Bereich „Schreibfeedback“ entnommen, in dem Schüler_innen Rückmeldungen zu ihren Aufsätzen anfordern. Die Aufsätze wurden manuell überprüft, um einen argumentativen Schreibstil zu gewährleisten. Insgesamt enthält das Korpus 7,116 Sätze mit 147,271 Wörtern.

Um eine hohe Datenqualität zu gewährleisten, wurde eine 31 Seiten umfassende Annotationsrichtlinie erarbeitet. Basierend auf dieser Richtlinie, annotierten drei Wissenschaftler aus dem Bereich der Sprachtechnologie unabhängig voneinander eine zufällige Auswahl von 80 Aufsätzen. Die Übereinstimmung zwischen den Annotatoren wurde mit Interrater-Reliabilität-Metriken systematisch evaluiert. Die Ergebnisse zeigten eine substantielle Übereinstimmung bei der Annotation von Argumentkomponenten. Unter Berücksichtigung der Komponententypen und deren Grenzen im Text erreichten die Annotatoren eine Übereinstimmung von $\alpha_U = .767$.[5] Die gemessene Interrater-Reliabilität der

[4] `www.essaysforum.com`

[5] $\alpha_U$ ist ein Interrater-Reliabilität-Metrik die sowohl die Kategorie (Kernaussage, Behauptung und Prämisse) als auch die Grenzen der markierten Textabschnitte auf Wortebene berücksichtigt.

argumentativen Relationen beträgt $\kappa = .708$ für den Relationstyp `belegt` und $\kappa = .737$ für den Relationstyp `widerlegt`. Diese Ergebnisse zeigen, dass das vorgeschlagene Annotationsschema mit hoher Übereinstimmung auf argumentative Aufsätze anwendbar ist und es möglich ist Trainingsdaten mit hoher Qualität zu erstellen. Ausgehend von diesen Ergebnissen wurde das komplette Korpus annotiert, womit der wissenschaftlichen Gemeinschaft erstmals Trainingsdaten für die ganzheitliche Erkennung von Argumentstrukturen in argumentative Aufsätzen zur Verfügung stand [SG14a, SG17a].

## 3 Automatische Erkennung von Argumentationsstrukturen

Die automatische Erkennung von Argumentationsstrukturen besteht aus mehreren Teilaufgaben. Diese umfassen die Trennung von argumentativen und nicht-argumentativen Textabschnitten und die Erkennung der Komponentengrenzen auf Wortebene (Segmentierung), die Klassifikation der argumentativen Rolle einer Komponente (z.b. als Behauptung oder Prämisse), und die Verlinkung der Argumentkomponenten mit argumentativen Relationen. Die Klassifikation von Argumentkomponenten und die Verlinkung von Argumentkomponenten sind jedoch nicht unabhängig voneinander [SG14b]. Insbesondere lässt sich die Funktion einer Argumentkomponente nur unter Berücksichtigung der argumentativen Diskursstruktur eines Textes bestimmen und umgekehrt. Um diese Herausforderung zu lösen und die wechselseitigen Informationen zu nutzen, wurde ein „Joint-Model“ entwickelt, das die Komponententypen und argumentativen Relationen global optimiert.

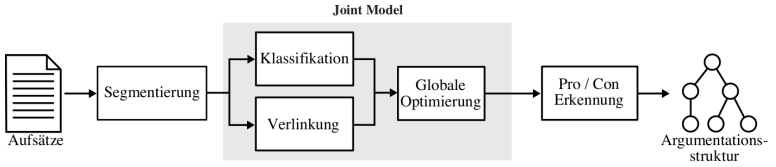



Abb. 2: Komponenten des Modells für die automatische Erkennung von Argumentationsstrukturen.

Abbildung 2 zeigt die Architektur des Modells mit den folgenden Komponenten:

1. *Segmentierung*: Die Segmentierung trennt argumentative von nicht-argumentativen Textabschnitten und erkennt die Grenzen von Argumentkomponenten auf Wortebene mit einem IOB-Schema[6] und einem *Conditional Random Field*.
2. *Joint-Model*: Das Joint-Model baut auf zwei lokalen Klassifikatoren auf, um die Funktion von Argumentkomponenten zu bestimmen und sie mit gerichteten argumentativen Relationen zu verknüpfen. Beide Modelle basieren auf einer *Support Vector Machine* (SVM). Die Ergebnisse beider Klassifikatoren werden anschließend global mit *Integer Linear Programming* (ILP) optimiert, um einen oder mehrere Bäume in jedem Paragraphen des Textes zu finden.

[6] Jedes Wort wird entweder als Beginn (B), innerhalb (I) oder außerhalb (O) einer Argumentkomponente gekennzeichnet.

3. *Pro/Con Erkennung*: Dieses Modell klassifiziert mit einer SVM jede Behauptung und Prämisse als "widerlegt" oder "belegt", um zwischen verschiedenen argumentativen Relationen zu unterscheiden.

Für jedes dieser Modelle wurden spezifische linguistische Eigenschaften definiert, um optimale Ergebnisse zu erzielen. Die Auswahl dieser Eigenschaften wurde für jedes Modell systematisch durch eine Kreuzvalidierung auf den Trainingsdaten durchgeführt. Die Ergebnisse zeigten beispielsweise, dass strukturelle Eigenschaften (bspw. die Position eines Satzes) effektiv für die Trennung von argumentativen und nicht-argumentativen Textabschnitten genutzt werden können und lexikalisch-syntaktische Informationen für die Erkennung des ersten Wortes einer Argumentkomponente effektiv sind. Die genauen Details zu dieser Untersuchung sind in Kapitel 5 der Dissertation beschrieben [St17].

Die Ergebnisse der lokalen Klassifikatoren werden mit folgendem ILP-Modell global optimiert, um eine konsistente Baumstruktur zu erhalten. Für $n$ gegebene Argumentkomponenten ist die Zielfunktion des Modells definiert als:

$$\underset{x,b}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_{ij} + b_{ij}, \tag{1}$$

wobei die Variablen $x_{ij} \in \{0,1\}$ eine argumentative Relation zwischen den Argumentkomponenten $i$ und $j$, und $b_{ij} \in \{0,1\}$ einen direkten Pfad von $i$ nach $j$ kennzeichnen.[7] Die Koeffizienten $w_{ij} \in \mathbb{R}$ sind Gewichte für jede Relation, die durch die Ergebnisse der lokalen Klassifikatoren bestimmt werden (siehe unten). Um eine Baumstruktur zu gewährleisten wird die Zielfunktion unter Berücksichtigung folgender Bedingungen gelöst:[8] Die erste Bedingung $\forall i : \sum_{j=1}^{n} x_{ij} \leq 1$ verhindert, dass eine Komponente mehr als eine ausgehende Kante besitzt. Um sicherzustellen dass mindestens eine Argumentkomponente ohne ausgehende Relation existiert (Wurzelknoten), wird die Anzahl aller Relationen durch die Bedingung $\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \leq n-1$ beschränkt. Die dritte Bedingung $\forall i : x_{ii} = 0$ verhindert dass eine Relation denselben Start- und Endknoten hat. Zusätzlich verhindern die drei folgenden Bedingungen (übernommen von [Kü08, p. 92]) Zyklen in der finalen Struktur. Die Bedingung $\forall i \forall j : x_{ij} - b_{ij} \leq 0$ koppelt die Hilfsvariablen $b_{ij}$ mit den Variablen $x_{ij}$ und gewährleistet, dass der Pfad zwischen den Argumentkomponenten $i$ und $j$ in $b_{ij}$ gesetzt ist, wenn es eine direkte Relation zwischen beiden Argumentkomponenten gibt. Mit der Bedingung $\forall i \forall j \forall k : b_{ik} - b_{ij} - b_{jk} \leq -1$ werden alle Pfade deren Länge größer als eins sind an die Hilfsvariablen gekoppelt. Durch die letzte Bedingung $\forall i : b_{ii} = 0$ werden alle Pfade die in derselben Argumentkomponente beginnen und enden – und somit Zyklen – vermieden.

Um die Ergebnisse der beiden lokalen Klassifikatoren in die Zielfunktion zu integrieren, werden die Gewichte für jede Relation als $w_{ij} = \phi_r r_{ij} + \phi_c c_{ij} + \phi_{cr} cr_{ij}$ definiert. Hierbei ist $r_{ij} = 1$, wenn eine argumentative Relation zwischen den Argumentkomponenten $i$ und $j$ gefunden wurde und 0 andernfalls. Die Variable $c_{ij}$ ist 1 falls Komponente $j$ als

[7] Jede Variable wird beim Lösen der Funktion als binär betrachtet.

[8] Zum Lösen verwenden wir das Framework lpsolve: `http://lpsolve.sourceforge.net`

Behauptung klassifiziert wurde, sonst ist $c_{ij} = 0$. Die dritte Komponente $cr_{ij} = cs_j - cs_i$ weißt Relationen von Behauptungen zu Prämissen einen höheren Wert zu. Der Wert $cs_i = \frac{relin_i - relout_i + n - 1}{rel + n - 1}$ ist die Wahrscheinlichkeit, dass Argumentkomponente $i$ eine Behauptung ist und wird durch die lokal gefunden Relationen bestimmt. Dabei ist $relin_i = \sum_{k=1}^{n} r_{ki}[i \neq k]$ die Anzahl der eingehenden Relationen von $i$, $relout_i = \sum_{l=1}^{n} r_{il}[i \neq l]$ die Anzahl der ausgehenden Relationen von $i$, und $rel = \sum_{k=1}^{n}\sum_{l=1}^{n} r_{kl}[k \neq l]$ die Anzahl der gefunden Relationen. Der Wert von $cs_i$ ist für eine Argumentkomponente mit vielen eingehenden Relationen größer als für eine Argumentkomponente mit weniger eingehenden Relationen. Jedes $\phi$ ist ein Parameter des ILP-Modells, um den Anteil der drei Informationen $r_{ij}$, $c_{ij}$, and $cr_{ij}$ zu steuern. Diese Parameter wurden anhand von Kreuzvalidierung bestimmt ($\phi_r = \frac{1}{2}$ und $\phi_{cr} = \phi_c = \frac{1}{4}$). Nach der Anwendung des ILP-Modells, werden die Relationen und Komponententypen entsprechend der resultierenden $x_{ij}$ gesetzt. Jede Komponente ohne ausgehende Kante wird als Behauptung und alle weiteren als Prämisse markiert.

Um das vorgeschlagene Modell zu evaluieren, wurde das annotierte Korpus in Trainings- (80%) und Testdaten (20%) aufgeteilt. Alle Parameter des Modells wurden durch systematische Kreuzvalidierung auf den Trainingsdaten bestimmt bevor die Modelle auf den Testdaten mit den folgenden zwei Baselines verglichen wurden: Die erste Baseline klassifiziert jede Instanz mit der häufigsten Klasse in den Daten (Mehrheits-Baseline). Die zweite Baseline (Heuristische Baseline) ist eine regelbasiertes System, welches anhand von Schreibrichtlinien für argumentative Aufsätze definiert wurde.[9] Tabelle 1 zeigt die

| | Segmentierung | Klassifikation | Verlinkung | Pro/Con | ∅ F1 |
|---|---|---|---|---|---|
| Mehrheits-Baseline | 25,9 | 26,0 | 45,5 | 47,8 | 36,3 |
| Heuristische Baseline | 64,2 | 75,9 | 70,0 | 56,2 | 66,6 |
| Lokale Klassifikatoren | † **86,7** | 79,4 | †71,7 | † **68,0** | 76,5 |
| ILP Joint Model | - | †‡ **82,6** | † **75,1** | - | **78,1** |

Tab. 1: Evaluationsergebnisse (Makro-F1-Metrik) der Argumenterkennung († = signifikante Verbesserung über der heuristischen Baseline; ‡ = signifikante Verbesserung über den lokalen Klassifikatoren; Wilcoxon-Test mit $\alpha = .005$).

Ergebnisse der Evaluation auf den Testdaten. Wie die Ergebnisse zeigen, übertrifft das Segmentierungsmodell die heuristische Baseline signifikant ($p = 1,65 \times 10^{-14}$). Die Evaluation des ILP-Modells zeigt, dass es erfolgreich die Ergebnisse der lokalen Klassifikation von Argumentkomponenten verbessert ($p = 7,45 \times 10^{-4}$) ohne die Ergebnisse des Modells zur Verlinkung negativ zu beeinflussen. Es erzielt 82,6 F1 für die Klassifikation von Argumentkomponenten und 75,1 F1 für die Erkennung von argumentativen Relationen. Eine genauere Untersuchung zeigt, dass das Modell die Erkennung von Behauptungen um 7,1 F1 und auch die Erkennung von verlinkten Argumentkomponenten um 7,7 F1 gegenüber den lokalen Modellen verbessert. Das Modell zur Unterscheidung von belegenden und widerlegenden Relationen verbessert die Klassifikationsgenauigkeit der heuristischen Baseline um 11,8 F1 ($p = 0.008$). Insgesamt zeigen die Ergebnisse, dass das ILP-Modell gleichzeitig die Klassifikation und die Verlinkung von Argumentkomponenten verbessert und

[9] Die Details zu dieser Baseline sind in [SG17a] beschrieben.

somit eine vielversprechende Grundlage für die Umsetzung von argumentativen Schreibsystemen bietet.

## 4 Automatische Qualitätsbewertung von Argumenten

Für die automatische Evaluation der Argumentqualität wurde ein Ansatz basierend auf dem Hinlänglichkeitskriterium von [JB77] entwickelt (engl. „sufficiency criterion“).[10] Ein Argument erfüllt dieses Kriterium, wenn dessen Prämissen ausreichen, um die Behauptung zu belegen. Das folgende Argument zum Thema Auslandsstudium veranschaulicht eine Verletzung des Hinlänglichkeitskriterium:

> **Beispiel 2)**: „*Einer meiner Freunde studierte Informatik an der Universität London und hat heute eine gut bezahlte Anstellung bei Google. Das zeigt, dass Studenten die im Ausland studierten besser bezahlt werden.*“

Die Prämisse dieses Beispiels bezieht sich auf ein einziges Beispiel (erster Satz), das die generelle Behauptung des Arguments belegen soll. Ein einziges Beispiel ist jedoch nicht ausreichend, um den generellen Fall zu bestätigen. Diese Art von Argument ist auch als voreilige Schlussfolgerung (engl. „*hasty generalization fallacy*“) bekannt [Go10].

> **Beispiel 3)**: „*Laut einer Umfrage verdienen Studenten die im Ausland studierten durchschnittlich 25% mehr pro Jahr. Daher könnte sich ein Auslandsstudium positiv auf das spätere Gehalt auswirken.*“

Die Behauptung in Beispiel 3 ist ausreichend begründet. Ausgehend von den Studienergebnissen kann man davon ausgehen, dass sich ein Auslandsstudium positiv auf das Gehalt auswirken könnte.

Um die Anwendbarkeit des Kriteriums für natürlichsprachliche Argumente zu evaluieren, annotierten drei Annotatoren 433 Argumente des in Abschnitt 2 beschriebenen Korpus. Die Ergebnisse zeigten, dass die Entscheidung der Annotatoren für $91,07\%$ der Argumente übereinstimmte. Die hohen Werte der Interrater-Reliabilität-Metriken von $\kappa = .7672$ und $\alpha = .7672$ bestätigen die substantielle Übereinstimmung. Ausgehend von diesen Ergebnissen wurden alle 1029 Argumente des Korpus annotiert. Insgesamt sind 33,8% dieser Argumente unzureichend begründet.

Zur automatischen Erkennung von unzureichend begründeten Argumenten wurden mehrere Ansätze in einer wiederholten Kreuzvalidierung evaluiert.[11] Das erste Modell ist eine SVM, welches anhand manuell definierter Eigenschaften ein gegebenes Argument als „ausreichend“ oder „unzureichend“ klassifiziert [SG17b]. Das zweite Modell ist ein *Convolutional Neural Network* (CNN) mit zuvor trainierten Word Embeddings. Tabelle 2 zeigt die Ergebnisse und den Vergleich beider Systeme mit einer Mehrheits-Baseline und einer SVM mit binären lexikalische Eigenschaften (SVM-Bow). Die SVM mit manuell definierten Eigenschaften verbessert die Ergebnisse der beiden Baselines signifikant. Es erreicht

[10] Kapitel 2 der Dissertation enthält einen detaillierten Vergleich von Qualitätskriterien aus Logik und Philosophie.

[11] Insgesamt wurden pro Modell hundert Experimente durchgeführt (5-fold CV mit 20 Wiederholungen).

| | Genauigkeit | F1 | F1 Unzureichend | Precision | Recall |
|---|---|---|---|---|---|
| Mehrheits-Baseline | 66,2 | 39,8 | 0 | 0 | 0 |
| SVM-bow Baseline† | 78,5 | 75,5 | 66,1 | 70,9 | 62,4 |
| SVM †‡ | 79,8 | 77,0 | 68,1 | 73,1 | 64,1 |
| CNN †‡ | **84,3** | **82,7** | **77,0** | **76,2** | **78,4** |

Tab. 2: Evaluationsergebnisse der Qualitätsbewertung († = signifikante Verbesserung über der mehreitlichen Baseline; ‡ signifikante Verbesserung über SVM-bow; Wilcoxon-Test mit $\alpha = 0.05$)

eine Genauigkeit von 79,8% und 77,0 F1. Die besten Ergebnisse erzielt das neuronale Netz. Er erreicht 82,7 F1 und eine Genauigkeit von 84,3%. Die genauere Betrachtung der Ergebnisse zeigt außerdem, dass das neuronale Netz eine deutlich bessere Trefferquote (engl. „recall“) als die SVM erzielt und sich somit besser für den Einsatz in argumentativen Schreibsystemen eignet.

## 5 Zusammenfassung

In dieser Dissertation wurden neue Ansätze zur automatischen Erkennung und Bewertung von natürlichsprachlichen Argumenten vorgestellt. Um die erforderlichen Korpora zu erstellen, wurden existierende Argumentationstheorien verglichen und ein Argumentationsmodell vorgestellt, welches die gesamte Argumentationsstruktur eines Dokuments als Baum modelliert. Wir zeigten erstmalig, dass menschliche Annotatoren Argumentationsstrukturen in argumentative Aufsätzen mit hoher Übereinstimmung identifizieren. Somit leistet diese Dissertation einen Beitrag für ein komplexes Problem der automatischen Sprachverarbeitung und dem jungen Forschungsbereich „Argument Mining“. Das Ergebnis dieser Annotationsstudie ist ein mit Argumentationsstrukturen annotiertes Korpus, welches von der Forschungsgemeinschaft vielfältig für weitere Arbeiten genutzt wird. Darüber hinaus stellten wir einen neuen Ansatz zur automatischen Erkennung von Argumentationsstrukturen vor. Dieser Ansatz erkennt die Grenzen von Argumentkomponenten auf Wortebene und optimiert die Funktion von Argumentkomponenten und argumentativen Relationen. Die Evaluationsergebnisse zeigen, dass dieser Ansatz nicht nur konsistente Argumentationsstrukturen erkennt, sondern im Vergleich zu mehreren Baselines auch signifikant bessere Erkennungsraten erzielt. Zusätzlich wurde ein Ansatz zur automatischen Qualitätsbewertung von Argumenten vorgestellt. Wir untersuchten erstmals die Eigenschaften von unzureichend begründeten Argumenten und stellten einen Ansatz basierend auf neuronalen Netzen vor, welcher im Vergleich mit mehreren Baselinesystemen signifikant bessere Ergebnisse erzielt.

## Literaturverzeichnis

[BB11] Butler, Jodie A.; Britt, M. Anne Britt: Investigating Instruction for Improving Revision of Argumentative Essays. Written Communication, 28(1):70–96, 2011.

[BMB10] Bentahar, Jamal; Moulin, Bernard; Bélanger, Micheline: A taxonomy of argumentation models used for knowledge representation. Artifi. Intelli. Rev., 33(3):211–259, 2010.

[Go10] Govier, Trudy: A Practical Study of Argument. Wadsworth, Cengage Learning, 7th. Auflage, 2010.

[JB77] Johnson, Ralph H.; Blair, Anthony J.: Logical Self-Defense. McGraw-Hill Ryerson, 1977.

[Kü08] Kübler, Sandra; McDonald, Ryan; Nivre, Joakim; Hirst, Graeme: Dependency Parsing. Morgan and Claypool Publishers, 2008.

[SB13] Shermis, Mark D.; Burstein, Jill: Handbook of Automated Essay Evaluation: Current Applications and New Directions. Routledge Chapman & Hall, 2013.

[SG14a] Stab, Christian; Gurevych, Iryna: Annotating Argument Components and Relations in Persuasive Essays. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014). S. 1501–1510, 2014.

[SG14b] Stab, Christian; Gurevych, Iryna: Identifying Argumentative Discourse Structures in Persuasive Essays. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). S. 46–56, 2014.

[SG17a] Stab, Christian; Gurevych, Iryna: Parsing Argumentation Structures in Persuasive Essays. Computational Linguistics, 43(3):619–659, 2017.

[SG17b] Stab, Christian; Gurevych, Iryna: Recognizing Insufficiently Supported Arguments in Argumentative Essays. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. S. 980–990, 2017.

[SN13] Schiappa, Edward; Nordin, John P.: Keeping Faith with Reason: A Theory of Practical Reason. Pearson Learning Solutions, 2013.

[St17] Stab, Christian: Argumentative Writing Support by means of Natural Language Processing. Dissertation, Technische Universität Darmstadt, 2017.

[Th73] Thomas, Stephen N.: Practical reasoning in natural language. Prentice-Hall, 1973.

[To58] Toulmin, Stephen E.: The Uses of Argument. Cambridge University Press, 1958.

[WRM08] Walton, Douglas; Reed, Chris; Macagno, Fabrizio: Argumentation Schemes. Cambridge University Press, 2008.

**Christian Stab** studierte Informatik an der Technischen Universität Darmstadt. Im Jahr 2009 schrieb er seine Diplomarbeit zum Thema „Interaktionsanalyse für adaptive Benutzerschnittstellen“. Nach seinem Studium arbeitete er vier Jahre im Bereich Informationsvisualisierung am Fraunhofer Institut für Graphische Datenverarbeitung. Im Oktober 2013 began er mit seiner Dissertation am UKP-Lab der Technischen Universität Darmstadt. Seine Dissertation schloss er im Februar 2017 mit Auszeichnung ab. Er ist (Ko)-Autor von mehr als dreißig Publikationen auf renommierten, internationalen Konferenzen. Seine Forschung im Bereich „Argument Mining“ wurde im Jahr 2016 mit einem „IBM Ph.D. Fellowship Award“ ausgezeichnet. Derzeit ist er Post-Doc am UKP-Lab und koordiniert das Validierungsvorhaben „ArgumenText“, in dem die neuesten Methoden zur automatischen Argumentextraktion in industrienahen Anwendungen erprobt werden.

# Face2Face: Übertragung von Gesichtsausdrücken in Echtzeit[1]

Justus Philipp-Andrei Thies[2]

**Abstract:**

Die Dissertation "Face2Face: Übertragung von Gesichtsausdrücken in Echtzeit" zeigt die Fortschritte der 3D Rekonstruktion von menschlichen Gesichtern und deren Anwendungen. Dabei wird darauf geachtet, dass die entwickelten Methoden mit herkömmlicher Endverbraucher-Hardware arbeiten. Durch diese Ansprüche an die Algorithmen, wird eine breite Anwendbarkeit sichergestellt, da keine komplizierten und teuren Aufbauten, wie sie zur Zeit in der Filmindustrie verwendet wird, nötig sind. Neben der 3D Erfassung des Gesichtes werden die Gesichtszüge auch über die Zeit verfolgt; dies geschieht in Echtzeit. Die Echtzeitkomponente ermöglicht zahlreiche neue Anwendungen, die in diversen Bereichen eingesetzt werden kann. Auf der einen Seite kann dadurch die Mensch-Maschinen Interaktion verbessert werden, in dem die Gesichtsform und Gesichtszüge in Echtzeit analysiert werden. Auf der anderen Seite stehen Anwendungen, die die Gesichtzüge auf virtuelle Avatare übertragen, wie dies auch aus der Filmindustrie zur Produktion von Animationsfilmen bekannt ist. Da die hier gezeigten Algorithmen zur digitalen Rekonstruktion des Gesichts auf dem Prinzip der Analyse durch Synthese beruhen, können dadurch auch virtuelle Avatare generiert werden. Diese Avatare können dann beliebig bearbeitet werden. Z.B. kann die Mimik verändert werden. Das sogenannte "Facial Reenactment" setzt genau diese Möglichkeit um. Hierzu werden die Gesichtausdrücke von einer Person auf einen virtuellen Avatar einer anderen Person übertragen. In dieser Arbeit wird das "Real-time Facial Reenactment" eingeführt und deren Anwendungen und Risiken aufgezeigt.

## 1 Einführung

Heutzutage sind Computer, Smartphones und Tablets aus unserem Alltag nicht mehr wegzudenken, sie sind all gegenwärtig. Damit diese Geräte mit der Umwelt interagieren können sind sie mit zahlreichen Sensoren ausgestattet. Kameras, Bewegungssensoren, Fingerabdruckscanner und viele mehr sind in modernen Geräten enthalten. Viele dieser Sensoren werden hauptsächlich dazu verwendet, um die Mensch-Maschinen Interaktion zu verbessern. Beispielsweise ersetzt ein Fingerabdrucksensor in vielen Bereichen das Passwort als Identifikationsmerkmal. Gleiches kann auch über Gesichtserkennung mit Hilfe einer Kamera erreicht werden. Eine Kamera kann aber auch in vielen weiteren Bereichen genutzt werden. Z.B. kann über eine Kamera die Augenbewegung des Nutzers verfolgt werden. Dadurch kann das Nutzerverhalten und der Blickfokus der Person analysiert werden. Beim sogenannten "Foveated Rendering" [Gu12], wird die Information genutzt um den Bereich, der vom Nutzer fokussiert wird, mit höheren Details zu zeichnen. Diese Technik nutzt dabei die Anatomie des menschlichen Auges aus (Auflösungsvermögen im peripheren Sehen ist geringer) und spart damit Rechenleistung ein.

[1] Englischer Titel der Dissertation: "Face2Face: Real-Time Facial Reenactment"

[2] Visual Computing Group, Technische Universität München, justus.thies@tum.de

Wie oben beschrieben kann eine Kamera wie ein Fingerabdruck für die Identifizierung eines Gesichtes genutzt werden. Neben dieser Identifizierung gibt ein Gesicht viele weitere Information über eine Person frei. Das Gesicht sagt viel über den Gefühlszustand einer Person aus [Ek82]. Für Menschen ist es ein leichtes diese Gefühlszustände wie Wut, Freude, etc. anhand eines Gesichtes abzuschätzen. Aktuelle Forschung versucht genau dies auch in Algorithmen abzubilden. Diese Informationen können dann auch genutzt werden um die Mensch-Maschinen Interaktion zu verbessern. Als Beispiel kann man hier an die Selektion von Musik je nach Gefühlslage denken. Es können aber auch Hinweisen entsprechend angepasst werden, z.B. Häufigkeit von Meldungen reduzieren, damit der Nutzer nicht gestört wird.

Neben der Analyse von Nutzerverhalten, werden Kameras auch zur Erfassung der Umgebung genutzt. Autonome Fahrzeuge sind auf die dreidimensionale Rekonstruktion der Umwelt angewiesen, um somit ihren Fahrweg zu planen. Eine dreidimensionale Rekonstruktion kann aber auch für diverse andere Planungs- oder Messaufgaben genutzt werden. Insbesondere werden für solche Zwecke spezielle Kameras benutzt, die neben Farb- auch Tiefendaten aufzeichnen. In Projekten wie ”KinectFusion” [Ne11] oder ”Real-time 3D Reconstruction at Scale Using Voxel Hashing” [Ni13] wird gezeigt wie aus solchen Kameraaufnahmen eine dreidimensionale Rekonstruktion von statischen Szenen erzeugt werden kann. Die abgescannten Objekte können dann auch mit Hilfe von 3D-Druckern dupliziert werden. Ein weites Anwendungsfeld haben solche Projekte auch im Bereich der Augmented Reality (AR) und Virtual Reality (VR). Objekte können ohne Zollstock vermessen werden und Modifikationen können virtuell simuliert werden. Virtuelle Spiegel, die verschiedene Make-up Vorschläge auf einem Gesicht simulieren können [Sc11] werden möglich. Man kann aber auch Kleidungsstücke passgenau für eine Person fertigen, deren dreidimensionales Körpermodell rekonstruiert wurde. Solche Zukunftsvisionen werden bereits von Unternehmen in Pilotprojekten umgesetzt (z.B. ”Adidas Knit for You” bei dem der Lehrstuhl für Graphische Datenverarbeitung in Erlangen beteiligt war).

All diese Anwendungen haben gemein, dass sie eine möglichst gute (und eventuell dynamische) dreidimensionale Rekonstruktion benötigen. In dieser Dissertation wird das Problem der Rekonstruktion von Gesichtern und der Verfolgung von Gesichtszügen behandelt. Um die Effektivität der entwickelten Verfahren zu demonstrieren, zeigen wir nicht nur Resultate der Rekonstruktion, sondern synthetisieren auch nahezu Photo-realistische Bilder von manipulierten Gesichtszügen. Dies erlaubt uns nicht nur die Umsetzung eines virtuellen Spiegels, sondern auch das Übertragen von Gesichtszügen auf eine andere Person, dem sogenannten ”Facial Reenactment”. D.h. man ist in der Lage das Gesicht einer Person virtuell zu steuern. Um dies umsetzen zu können, wird das Gesicht der zwei involvierten Person rekonstruiert und verfolgt. Dabei unterscheiden wir zwischen der Quellperson und der Zielperson. Die Gesichtsausdrücke der Quellperson werden extrahiert und auf das rekonstruierte Gesicht der Zielperson übertragen. Diese Manipulation geschieht zunächst im dreidimensionalen und wird anschließend mit Methoden aus der Computergraphik zu einem zweidimensionalen Bild mit der entsprechenden Mimik der Quellperson, welches über das original Zielbild gelegt wird.

Die vorgestellten Verfahren zur Erfassung der Gesichtszüge und deren Übertragung hat einige Anwendungsbereiche. In der Filmproduktion kann es als Editierwerkzeug eingesetzt werden, um die Mimik eines Schauspielers nachträglich zu verändern. Es können aber auch Einstellungen wie Beleuchtung oder Make-up angepasst werden. Eines der wichtigsten möglichen Anwendungsgebiete in der Postproduktion ist jedoch die Nachvertonung. Also die Synchronisierung von Mundbewegung zu der Stimme des Synchronsprechers. Heutzutage wird bei der Übersetzung eines Films in eine andere Sprache oft der Text so angepasst, dass er zum Videomaterial passt. Dadurch können Details verloren gehen. Aber in vielen Fällen ist trotz dieser Maßnahmen Ton und Video nicht synchron. Mit unseren Techniken kann die Mimik und Mundbewegung des Dolmetschers direkt auf eine Zielperson übertragen werden. Dadurch wird sichergestellt, dass Bild und Ton zueinander passen. Da wir die entwickelten Algorithmen für den Echtzeiteinsatz erschaffen haben, kann diese Dolmetscheranwendung auch für Live-Telekonferenzen genutzt werden, in der eine Übersetzung von Nöten ist (→ Simultandolmetscher).

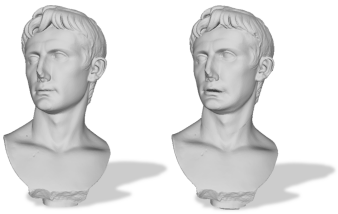


Abb. 1: Virtuelle Augustus Büste: Links statischer 3D Scan, rechts animierte Büste.

Im Gegensatz zu herkömmlichen Systemen zur Rekonstruktion von Gesichtszügen in der Filmindustrie, die mit speziellen Markern und komplexen Kameraaufbauten arbeiten, benötigen wir ein einfaches Setup welches aus einer einzigen Kamera besteht und Marker-los operiert. Unsere Erfassung der Mimik kann auch benutzt werden, um virtuellen Charakteren/Avataren Leben einzuhauchen und sie zu animieren (siehe z.B. Abb. 1). Neben Zeichentrickfilmen, werden solche virtuellen Charaktere in Computerspielen genutzt. Mit der Einführung der VR-Brillen, sind realistische Animationen solcher virtuellen Avatare immer wichtiger geworden, damit das Spielerlebnis immersiv ist. Das *FaceVR* Projekt (siehe Abschnitt 5) zeigt, dass die Rekonstruktion der Gesichtszüge auch möglich ist, wenn der Nutzer eine VR-Brille trägt, die annähernd die Hälfte des Gesichtes verdeckt. Das Projekt ebnet dabei auch den Weg zur Telekonferenz in VR, bei der die VR-Brille virtuell entfernt wird.

Neben dieser Verbraucheranwendungen, zeigen Forscher in der Psychologie ein starkes Interesse an unseren Projekten. Dabei möchten sie untersuchen, welchen Einfluss das Gesicht einer Person während eines Gesprächs auf einen Menschen hat. U.a. wollen die Forscher auch herauszufinden, ob das Gesicht einen Einfluss auf die Glaubwürdigkeit einer Aussage hat.

Die ursprüngliche Idee der Rekonstruktion von Gesichtern, stammt von einem medizinischen Projekt, bei dem Patienten die an einer Mund-Kiefer-Gaumenspalte litten untersucht wurden. Über einen längeren Zeitraum wurden dazu Bilder von Patienten aufgenommen und somit der Heilungsprozess dokumentiert. Ziel dieses Projektes war es eine Aussage für neue Patienten zu treffen und deren Heilungsprozess zu simulieren. Man kann sich mit unseren Forschungsergebnissen auch ein Trainingssystem für Patienten vorstellen, die einen Schlaganfall erlitten haben und nun ihre Gesichtsmimik trainieren müssen. Um Operationen zu planen, sind dreidimensionale Erfassungen in der modernen Medizin unab-

kömmlich. Aber nicht nur für die Planung sondern auch für die Durchführung kann ein Rekonstruktion eines Kopfes eingesetzt werden. So kann durch das Verfolgen des Kopfes während einer Operation zusätzliche Informationen, wie z.b. CT Daten durch AR Techniken eingeblendet werden.

Unsere Rekonstruktion und die Fähigkeit der Synthetisierung neuer Bilder macht die Manipulation von beliebigen Videos in Echtzeit möglich. In Kombination mit einem Sprachimitator oder einer Personen-spezifischen Stimmsynthese, kann man gefälschte Videos erzeugen, die für bösartige Zwecke wie Diffamierung, Fälschung von Beweisen oder Propaganda eingesetzt werden. Diese auch unter dem Stichwort ”Fake-News” entfachte Diskussion über die Glaubwürdigkeit von Video-,Bild- und Tonbeweisen ist äußerst relevant. Hierbei muss jedoch gesagt werden, dass Manipulationen, wie die von unserem *Face2Face* Projekt erstellt werden können, auch vorher mit entsprechendem Aufwand möglich waren / sind. Mit unserem offenen Umgang mit den Forschungsergebnissen und den zahlreichen öffentlichen Vorführungen hoffen wir, dass wir einen Teil zu dieser Diskussion beitragen konnten und die Menschen zu einer Neueinschätzung des Wertes von Videomaterial aus unbekannter Herkunft zu bringen. Die digitale Forensik und die Erkennung von Manipulationen gewinnt heutzutage immer mehr an Wichtigkeit. Als Nebeneffekt unserer Rekonstruktion kann die physikalische Plausibilität eines Bildes überprüft werden. Ein wichtiger Indikator ist dabei die Beleuchtung. Ist die Beleuchtung im Gesicht nicht konsistent zu der Beleuchtung eines anderen Objektes in der Szene kann von einer Fälschung ausgegangen werden.

Zusammenfassend lässt sich sagen, dass unser Hauptziel das Erstellen von mathematischen Repräsentationen der realen Welt ist. Die dazu entwickelten Algorithmen ermöglichen es Computern die Welt zu rekonstruieren, zu verstehen und mit ihr zu interagieren. In dieser Dissertation konzentrierten wir uns auf die Erfassung von Gesichtern die sich dynamisch bewegen - auch in unkontrollierten Umgebungen.

## 2 Grundlagen

Die Grundlage der Rekonstruktion von Gesichtern bildet ein digitales Gesichtsmodell. Dieses Gesichtsmodell basiert auf den Daten von Blanz und Vetter [BV99], die 200 Gesichter digital erfassten (Form und Farbe; jeweils 100 männliche und weibliche Personen). Durch eine Hauptkomponentenanalyse kann sowohl das durchschnittliche Gesicht, als auch die Hauptkomponenten bestimmt werden. Rechts werden beispielsweise jeweils der Einfluss für die Gesichtsform als auch für die Gesichtsfarbe gezeigt. Wie man sieht, gibt die erste Hauptkomponente der Gesichtsform zum großem Teil die Gesichtsgröße an, blendet aber auch zwischen einem männlichen und einem weiblichen Gesicht. Die erste Hauptkomponente der Gesichtsfarbe blendet zwischen einem hellen und dunklen Hautton.

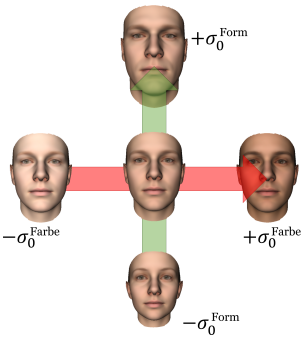



Abb. 2: Gesichtsmodell.

Mit Hilfe dieser Hauptkomponenten können durch ein Linearkombination neue Gesichter berechnet werden. Die Koeffizienten der Linearkombination repräsentieren dabei sozusagen die Identität der Person in einem niedrig dimensionalen Raum. Dieses statistische Gesichtsmodell kann jedoch nur das Gesicht in einer neutralen Mimik darstellen. Um Gesichtsausdrücke darstellen zu können, wurden daher sogenannte ”Blendshapes” für dieses statistische Gesichtsmodell generiert. ”Blendshapes” geben Beispielposen des Gesichtes

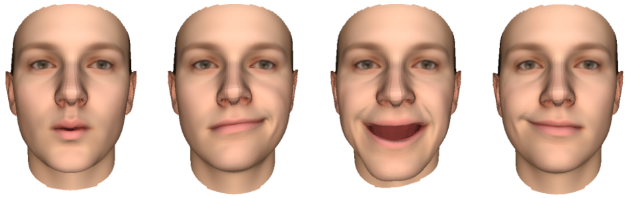


Abb. 3: Modellierung von Gesichtsausdücken durch Beispielposen sogenannten ”Blendshapes”.

an. Solche Posen kann man zum Beispiel in Abb. 3 sehen. Zwischen diesen Posen kann interpoliert werden, um beliebige Kombinationen zu erreichen. Um diese Beispielposen für beliebige Gesichtsformen zu generieren und nicht nur für das durchschnittliche Gesicht, werden Delta-Blendshapes eingesetzt. Dazu wird die Differenz der Beispielposen zum neutralen Gesicht berechnet. Diese Differenzvektoren können auf andere Gesichter angewendet werden. Ähnlich wie die Erstellung neuer Gesichter durch das statistische Gesichtsmodell, kann durch eine Linearkombination der Delta-Blendshapes die Mimik eines Gesichtes dargestellt werden - die nötigen Koeffizienten werden Blendshape-Gewichte genannt.

Durch die Identitätsparameter und die Blendshape-Gewichte kann somit ein Gesicht digital approximiert werden. Die automatische Berechnung dieser Modellparameter ist dabei eine der Hauptkomponenten für die Projekte, welche in den nachfolgenden Abschnitten gezeigt werden.

Als Grundlage der Berechnung der Modellparameter dient das Prinzip der *Analyse durch Synthese*. Dabei werden die Modellparameter optimiert, um ein oder mehrere Eingabebilder möglichst gut nachzubilden. Dazu muss neben der Gesichtsfarbe und Form des Gesichtes auch Beleuchtung, die rigide Position und Orientierung des Gesichts, und die Parameter eines Kameramodells geschätzt werden. Das Kameramodell beschreibt dabei den Bildgenerierungsprozess, d.h. die Abbildung vom dreidimensionalen Gesicht zu einem zweidimensionalen Bild. Um die Modellparameter anzupassen, muss eine Fehlermetrik definiert werden. Diese ist vorallem abhängig von der Art der Eingabebilder (z.B. Tiefenbilder oder Farbbilder), aber auch von der Komplexität (z.B. ob alle Bildpunkte berücksichtigt werden oder nur einzelne wenige). In der Dissertation werden mehrere solche Fehlermetriken behandelt. Eine Gemeinsamkeit ist dabei, dass alle Bildpunkte des synthetisierten Bildes mit dem originalen Eingabebild verglichen werden. Dies impliziert eine hohe Anzahl von Berechnungen, die jedoch mit modernen Grafikkarten effizient abgearbeitet werden können. Um den Fehler der Fehlermetrik zu minimieren, wird ebenfalls die Rechenleistung der Grafikkarte ausgenutzt. Dazu wurde ein speziell für das Problem entwickelter Gauss-Newton Ansatz implementiert. Für eine detaillierte Beschreibung der Optimierung und der Fehlermetriken sei auf die originale Dissertation verwiesen.

## 3 Echtzeit Übertragung von Gesichtsausdücken


Abb. 4: Echtzeit Übertragung von Gesichtsausdücken mit Hilfe von speziellen RGB-D Kameras.

Das initiale Projekt zur Echtzeitübertragung von Gesichtsausdrücken hat den englischen Namen „Real-time Expression Transfer for Facial Reenactment" [Th15]. In Abb. 4 wird das Szenario des Projektes gezeigt. Die Gesichtsausdrücke der Person rechts im Bild wird virtuell auf die linke Person übertragen; das Resultat wird in der Mitte auf den Bildschirmen gezeigt. Die Besonderheit an dem System ist die Übertragung und das Photo-realistische Neuzeichnen der Gesichtsausdrücke in einem Zielvideo. Um dies umzusetzen, werden die Gesichtszüge des Quellvideos rekonstruiert und verfolgt. Dazu greifen wir auf eine spezielle Farb- und Tiefenkamera zurück. Eine solche Kamera ist heutzutage in vielen Geräten zu finden. U.a. ist hierbei der Kinect Sensor von Micosoft zu erwähnen, da diese Kamera den Grundstein für Tiefensensoren im Verbrauchersektor gelegt hat. Ähnliche Sensoren sind nun auch in Laptops, Tablets und Smart-Phones zu finden. Basierend auf den Bilddaten einer solchen Kamera schätzen wir gleichzeitig die Modellparameter für Identität, Gesichtsausdrücken, rigide Pose und Beleuchtung. Dabei wird das Farb- und das Tiefenbild verwendet und eine Minimierung der Fehlerquadrate zwischen realen und synthetischen Bildern in jedem Zeitschritt in Echtzeit berechnet. Dies wird für beide Personen durchgeführt. Da die Optimierung der Modellparameter auf dem Prinzip der Analyse-durch-Synthese beruht, sind wir in der Lage nahezu Photo-realistische Bilder der Personen zu synthetisieren. Basierend auf den geschätzten Modellparametern können die Differenzen der Mimik zwischen Quell- und Zielperson im Parameterraum berechnet werden und die Ausdrücke der Zielperson entsprechend der Quellperson angepasst werden. In Abb. 5 wird der Ablauf gezeigt, wie aus diesen Informationen ein neues Bild generiert wird. Das Gesicht der Zielperson, welches dreidimensional rekonstruiert wurde, wird entsprechend der Mimik der Quellperson angepasst. Mit der geschätzten Beleuchtung im Zielbild kann das Gesicht neu gezeichnet werden. Da das parametrische Gesichtsmodell keinen Mundinnenraum modelliert, wird für die Synthese ein generisches Zahnmodell und eine Textur für den Mundinnenraum genutzt. Anschließend wird dieses Bild auf das Zielbild gelegt, dies ist möglich da wir die rigide Pose der Zielperson berechnet haben. Die Robustheit des Systems wird in dem Paper welches bei der Siggraph Asia 2015 präsentiert wurde und den dazugehörigen Videos gezeigt. Das Live-System wurde auch auf diversen Konferenzen Besuchern demonstriert, u.a. auf der GPU Technology Conference 2016.

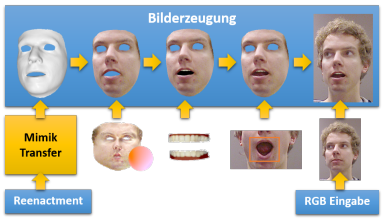



Abb. 5: Synthese des Ausgabebildes.

## 4 Echtzeit Manipulation von Videos

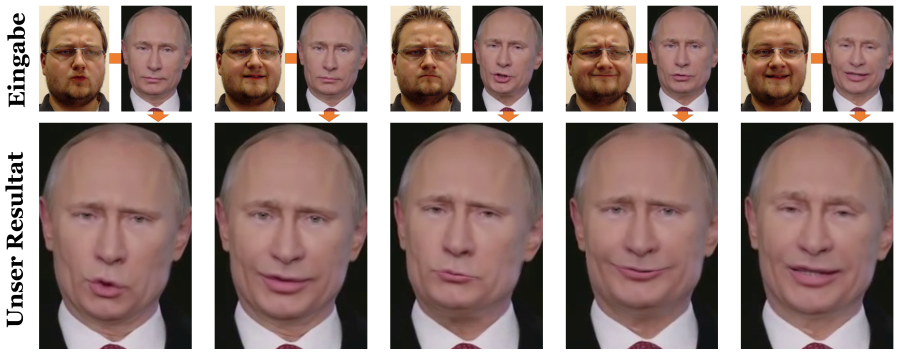



Abb. 6: Face2Face erlaubt es beliebige Videos aus dem Internet in Echtzeit zu bearbeiten.

Wie im vorangehenden Kapitel beschrieben, ist es möglich basierend auf speziellen Tiefenkameras Gesichter zu rekonstruieren und diese zu manipulieren. Obwohl solche Kameras bereits eine weite Verbreitung haben, sind Videos solcher Kameras sehr selten zu finden. Videos aus dem Internet oder Filme liegen meinst nur als gewöhnliche Farbbilder vor. Eine Rekonstruktion von Gesichtern anhand solche Bilder ist ungleich schwerer als die Rekonstruktion basierend auf Tiefendaten, da keine direkten Messdaten der Geometrie vorliegen. Im Projekt „Face2Face: Real-time Face Capture and Reenactment of RGB Videos" [Th16a], wird genau dieses Problem angegangen und in Echzeit gelöst. Zur Rekonstruktion wird dabei eine neues Minimierungsproblem beschrieben, welches auf einer robusten $\ell_{2,1}$-Norm beruht. Im Gegensatz zu den Tiefenkameras liegen bei Videos meist keine Kalibrierdaten vor. D.h. es müssen auch die Kameraparameter geschätzt werden. Um das Fehlen der Tiefe zu kompensieren, werden zur Rekonstruktion der Identität mehrere Bilder des Videos gleichzeitig benutzt. Alle verwendeten Bilder haben dabei gemein, dass die Identitätsparameter gleich sind, jedoch kann die Person in den einzelnen Bildern verschiedene Gesichtsausdrücke,rigide Posen und Beleuchtungen aufweisen. Nachdem die Identitätsparameter geschätzt sind, können die weiteren Parameter für folgende Bilder in Echtzeit geschätzt werden. Dieses Verfahren setzen wir sowohl für ein Quell-, als auch für ein Zielvideostream ein. Das Quellvideo kann dabei eine WebCam und das Zielvideo ein Video aus dem Internet sein (wie auch in Abb. 6 zu sehen). Abb. 7 beschreibt das Vorgehen. Das Video der Zielperson wird vorab analysiert, d.h. es werden für jedes Bild die Modellparameter geschätzt. Zur Laufzeit werden die Parameter der Quellperson berechnet und dann auf die andere Person übertragen. Bei der Übertragung der Mimik werden auch die Geometrien der beiden Personen berücksichtigt. Um ein möglichst realistisches Bild zu synthetisieren, extrahieren wir aus der Videosequenz der Zielperson Texturen des Mundinnenraums.Aus dieser Datenbank wird die Textur ausgewählt, welche am nächsten an dem gewünschten Gesichtsausdruck liegt und anschließend in das Zielbild projiziert.

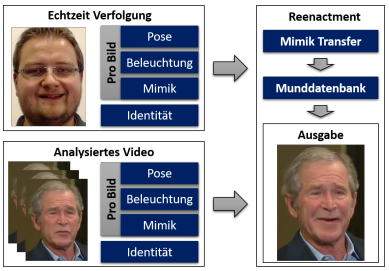



Abb. 7: Face2Face Überblick.

Das Face2Face Projekt wurde im Rahmen der CVPR 2016 vorgestellt und demonstriert. Außerdem zeigten wir eine Live-Demonstration bei der Siggraph Emerging Technologies 2016 [Th16c] und wurden mit dem ”Best in Show Award” ausgezeichnet.

## 5 Telekonferenzen in der Virtuelle Realität

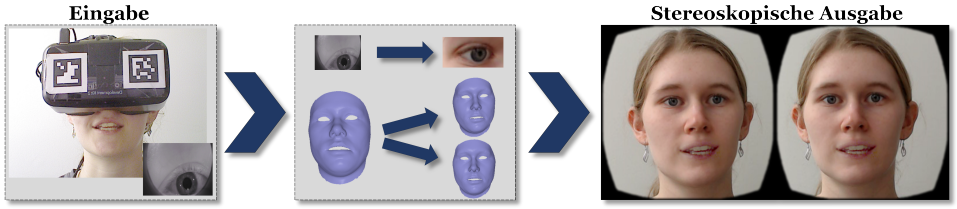



Abb. 8: Telekonferenzen in der Virtuelle Realität: FaceVR ermöglicht es virtuell die VR-Brille zu entfernen, was für eine VR-Telekonferenz unabdingbar ist, da ansonsten die Hälfte des Gesichtes des Gesprächspartners verdeckt ist.

In *FaceVR* [Th16b] nutzen wir die Möglichkeit der Mimikübertragung um virtuell eine Virtual Reality (VR) Brille zu entfernen. Dies wird zum Beispiel für eine Videokonferenz in VR benötigt, da ansonsten die Hälfte des Gesichts durch die Brille verdeckt wird (wie in Abb. 8 links zu sehen). Die Rekonstruktion der Modellparameter musste für dieses Szenario stark angepasst werden. Um die regide Pose des Kopfs der Quellperson robust zu verfolgen, benutzen wir ArUco AR Marker die auf die Brille geklebt wurden. Zusätzlich wurde in das Headset eine Infrarotkamera eingebaut, mit deren Hilfe die Augenbewegungen verfolgt werden können. Die Mimikparameter der Quellperson werden mit einer Tiefenkamera verfolgt, dabei beschränkt sich die Erfassung jedoch nur auf den nicht verdeckten Bereich des Gesichtes.

Da wir eine Videokonferenz in VR ermöglichen möchten, wurde das Zielvideo mit einer Stereokamera aufgezeichnet. Dadurch kann ein stereoskopisches Bild direkt auf einer VR-Brille angezeigt werden. Dieses Video wird dann mit einer Erweiterung des Face2Face Verfahrens analysiert. Hierbei werden jeweils das linke und rechte Bild der Kamera gleichzeitig in der Modellparameterbestimmung berücksichtigt. Durch diese zusätzlichen Daten können auch besser Rekonstruktionen erzielt werden, wie in der Publikation gezeigt wird. Zur Synthetisierung neuer Bilder, werden die Gesichtsausdrücke und die Augenbewegungen auf das Ziel-Stereovideo angewendet. Um möglichst realistische Augen zu zeichnen, benutzen wird eine Kalibriersequenz, in der die Zielperson in bestimmte Richtungen blicken muss. Ausgehend von diesen Beispielbildern und der Blickrichtung der Quellperson werden dann die Augen synthetisiert.

Dieses Projekt wurde wie die anderen Projekte der Öffentlichkeit im Rahmen der Siggraph 2017 Emerging Technologies gezeigt.

# 6 Zusammenfassung

In dieser Dissertation werden die Fortschritte im Bereich der 3D-Rekonstruktion von Gesichtern, basierend auf herkömmlicher Endverbraucher-Hardware gezeigt. Neben der Rekonstruktion der Geometrie und der Textur eines Gesichtes, wird auch die Verfolgung von Gesichtszügen in Echtzeit demonstriert. Die entwickelten Algorithmen basieren auf dem Prinzip der Analyse durch Synthese. Um dieses Prinzip anwenden zu können, muss zuerst ein mathematisches Modell definiert werden, welches es ermöglicht ein Gesicht virtuell darzustellen. Neben dem Gesichtsmodell wird auch der Aufnahmeprozess der verwendeten Kamera in einem Modell dargestellt. Durch die Möglichkeit ein Bild eines Gesichtes zu synthetisieren, können iterativ die Modellparameter so angepasst werden, dass das synthetisierte Bild bestmöglich das Eingabebild repräsentiert. Mit Hilfe dieses Verfahrens überführt man somit im Umkehrschluss das Eingabebild in eine virtuelle Darstellung eines Gesichtes. Die erreichte Qualität ermöglicht eine Vielzahl von neuen Anwendungen, die auf eine detailgetreue Rekonstruktion angewiesen sind. Dazu gehört auch das sogenannte ”Facial Reenactment”. Unsere entwickelten Methoden zeigen, dass eine solche Anwendung ohne spezielle Ausrüstung möglich ist. Die Resultate sind nahezu Photo-realistische Videos, in denen die Mimik einer Person auf eine andere Person übertragen wird. Dadurch lässt sich zum Beispiel die Synchronisierung von Filmen, also das Übersetzen in eine andere Sprache verbessern. Anstatt die Audiospur an das Video anzupassen, was unter anderem auch zu Änderungen am Text führt, können die Mundbewegungen des Dolmetschers in einem Nachbearbeitungsschritt des Videomaterials auf den Schauspieler übertragen werden. Da die Techniken, die in dieser Dissertation gezeigt werden, in Echtzeit ablaufen, kann auch in einem Videotelekonferenzsystem die Mundbewegung eines Live-Dolmetschers virtuell auf eine andere Person übertragen werden.

Die Veröffentlichungen der Videos zu denen in der Dissertation gezeigten Projekten, führten zu einer breiten Diskussion in den Medien. Dies lag zum einen an der Tatsache, dass unsere Methoden so entwickelt wurden, dass sie in Echtzeit ablaufen können und zum anderen daran, dass wir die Anforderungen an Hardware auf ein Minimum reduziert konnten. So ist es möglich, gewöhnliche Videos aus dem Internet zu bearbeiten und in Echtzeit zu editieren. Unter anderem haben wir somit bekannten Persönlichkeiten, wie zum Beispiel ehemaligen Präsidenten der USA, eine andere Mimik auferlegen können. Dies führte unweigerlich zu einer Diskussion über die Glaubwürdigkeit von Videomaterial, vor allem aus unbekannten Quellen. Das eine solche Manipulation bereits vor unseren gezeigten Demonstrationen möglich war, wenn auch mit einem höheren Aufwand, war den meisten Menschen nicht bewusst. Damit konnten wir mit unseren Projekten, neben der Weiterentwicklung von Echtzeit Gesichtsrekonstruktion, zu einer Sensibilisierung der Öffentlichkeit beitragen.

## Literaturverzeichnis

[BV99] Blanz, Volker; Vetter, Thomas: A morphable model for the synthesis of 3D faces. In: Proc. SIGGRAPH. ACM Press/Addison-Wesley Publishing Co., S. 187–194, 1999.

[Ek82] Ekman, P.: Emotion in the Human Face. Cambridge University Press, 1982.

[Gu12] Guenter, Brian; Finch, Mark; Drucker, Steven; Tan, Desney; Snyder, John: Foveated 3D Graphics. ACM SIGGRAPH Asia, November 2012.

[Ne11] Newcombe, Richard A.; Izadi, Shahram; Hilliges, Otmar; Molyneaux, David; Kim, David; Davison, Andrew J.; Kohli, Pushmeet; Shotton, Jamie; Hodges, Steve; Fitzgibbon, Andrew: KinectFusion: Real-time Dense Surface Mapping and Tracking. In: Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality. ISMAR '11, IEEE Computer Society, Washington, DC, USA, S. 127–136, 2011.

[Ni13] Nießner, Matthias; Zollhöfer, Michael; Izadi, Shahram; Stamminger, Marc: Real-time 3D Reconstruction at Scale Using Voxel Hashing. ACM Trans. Graph., 32(6):169:1–169:11, November 2013.

[Sc11] Scherbaum, Kristina; Ritschel, Tobias; Hullin, Matthias; Thormählen, Thorsten; Blanz, Volker; Seidel, Hans-Peter: Computer-suggested Facial Makeup. S. 485–492, 2011.

[Th15] Thies, J.; Zollhöfer, M.; Nießner, M.; Valgaerts, L.; Stamminger, M.; Theobalt, C.: Real-time Expression Transfer for Facial Reenactment. ACM Transactions on Graphics (TOG), 34(6), 2015.

[Th16a] Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; Nießner, M.: Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In: Proc. CVPR. S. 2387–2395, 2016.

[Th16b] Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; Nießner, M.: FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality. ArXiv, non-peer-reviewed prepublication by the authors, abs/1610.03151, 2016.

[Th16c] Thies, Justus; Zollhöfer, Michael; Stamminger, Marc; Theobalt, Christian; Nießner, Matthias: Demo of Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In: ACM SIGGRAPH 2016 Emerging Technologies. SIGGRAPH '16, ACM, New York, NY, USA, S. 5:1–5:2, 2016.

Justus Thies wurde am 18. Januar 1989 in Buchen (Odenwald) geboren. Nach seinem Abitur im Jahr 2008, begann er an der Friedrich-Alexander Universität Erlangen-Nürnberg sein Informatik Studium. Im November 2011 erlangte er dort sein Bachelor of Science, im Februar 2014 folgte der Master of Science mit Auszeichnung. Der Schwerpunkt seines Studiums lag dabei im Anwendungsorientierten Bereich, u.a. Computer Graphik, Medizinische Bildverarbeitung, Mustererkennung aber auch Digitale Forensik. Mit diesem Hintergrund begann Justus Thies im April 2014 sein Doktorrat unter der Betreuung von Prof. Dr. Günther Greiner. Im April 2017 reichte er seine Dissertation ein, die er schließlich im Oktober 2017 erfolgreich verteidigte. Seit September 2017 ist Justus Thies an der Technischen Universität München in der Visual Computing Group von Prof. Dr. Matthias Nießner tätig. Dort führt er seine Forschung weiter, u.a. behandelt er dabei auch die Digitale Forensik und das Erkennen von Bildmanipulationen die durch Techniken wie Face2Face erstellt wurden.

# Beiträge zu praktikabler Prädikatenanalyse[1]

Philipp Wendler[2]

**Abstract:** Der Stand der Forschung im Bereich der automatischen Software-Verifikation ist fragmentiert. Verschiedene Verfahren existieren nebeneinander in unterschiedlichen Darstellungen und mit wenig Bezug zueinander, aussagekräftige Vergleiche sind selten. Die Dissertation adressiert dieses Problem. Ein konfigurierbares und flexibles Rahmenwerk zur Vereinheitlichung solcher Verfahren wird entwickelt und mehrere vorhandene Verfahren werden in diesem Rahmenwerk ausgedrückt. Dies bringt neue Erkenntnisse über die Kernideen dieser Verfahren, ermöglicht experimentelle Studien in einer neuartigen Qualität, und erleichtert die Forschung an Kombinationen und Weiterentwicklungen dieser Verfahren. Die Implementierung dieses Rahmenwerks im erfolgreichen Verifizierer CPACHECKER wird in der bisher größten derartigen experimentellen Studie (120 verschiedene Konfigurationen, 671 280 Ausführungen) evaluiert. Hierzu wird ein Benchmarking-System präsentiert, das mit Hilfe moderner Technologien signifikante qualitative Messfehler existierender Systeme vermeidet.

## 1 Einführung

Software ist ein wichtiger Bestandteil unseres Lebens und steuert viele sicherheitskritische Systeme wie z. B. Kraftwerke, Flugzeuge, Züge und Autos. Nicht nur in solchen Systemen ist die Korrektheit der eingesetzten Software von höchster Wichtigkeit. Testen ist ein gängiges Verfahren zum Finden von Fehlern in Software, kann jedoch nicht alle Fehler finden bzw. deren Abwesenheit beweisen. Manuelle oder semi-automatische Verifikation ist aufgrund des Aufwands oft nicht einsetzbar. Automatische Verifikation versucht diese Lücke zu schließen und höhere Gewissheit als Testen mit weniger Aufwand als manuelle Verifikation zu erreichen: Gegeben ein Eingabeprogramm und eine Spezifikation soll ein Verifizierer automatisch entscheiden, ob das Programm die Spezifikation erfüllt oder nicht. Aufgrund der Unentscheidbarkeit des Problems kann es allerdings zu Fällen ohne nutzbarem Ergebnis kommen. Obwohl sich Verfahren wie Model-Checking z. B. in der Verifikation von Gerätetreibern in Betriebssystemen als nützlich erwiesen haben, werden sie von Software-Entwicklern allerdings nur selten eingesetzt.

Fortschritt im Bereich automatischer Software-Verifikation hängt ab von einem theoretischen Verständnis der Verfahren und der Möglichkeit, diese effektiv in der Praxis durch Benchmarking zu evaluieren. Hierbei können wir drei Probleme identifizieren:

1. Vorgestellte Verifikationsansätze unterscheiden sich oft grundlegend in ihrer Darstellung, selbst bei verwandten Ansätzen. Dies erschwert das Verständnis und die Möglichkeit, die Kernideen der verschiedenen Ansätze zu identifizieren und darauf aufbauend neue Ansätze zu entwickeln.

[1] Englischer Titel der Dissertation: „Towards Practical Predicate Analysis“

[2] Lehrstuhl für Software and Computational Systems, LMU München, philipp.wendler@lmu.de

2. Eine experimentelle Evaluation ist oft schwierig, da qualitativ hochwertige Implementierungen von Ansätzen nicht immer zur Verfügung stehen, und die vorhandenen Implementierungen in verschiedenen Verifizierern existieren, was aufgrund des Einflusses technischer Unterschiede die Vergleichbarkeit von Ergebnissen einschränkt.
3. Die zum Benchmarking im Rahmen von experimentellen Evaluationen eingesetzten Tools sind nicht immer zuverlässig und können Messfehler beliebiger Größe produzieren.

### 1.1 Zielsetzung

Ziel der Dissertation [We17] ist es, eine Lösung für diese Probleme zu finden. Wir fokussieren uns hierbei insbesondere auf Verfahren des Model-Checkings mit Hilfe von Prädikaten über Programmvariablen. Dies erlaubt es von der Mächtigkeit und Effizienz moderner Solver für Erfüllbarkeit modulo Theorien (SMT) zu profitieren. Bekannte prädikatenbasierte Verfahren sind z. B. Prädikatenabstraktion und Bounded Model-Checking, die in einer Reihe bekannter Verifizierer implementiert sind und in der Praxis eingesetzt werden.

In der Dissertation definieren wir ein flexibles theoretisches Rahmenwerk, in dem verschiedene bekannte Verfahren ausgedrückt und so vereinheitlicht werden. Dies ermöglicht es diese Verfahren zu vergleichen, ohne dass oberflächliche Darstellungsunterschiede dies erschweren, die jeweiligen Kernideen der Verfahren zu identifizieren, und neue, effektivere und effizientere, Kombinationen von Verfahren auf einfache Weise zu definieren. Außerdem entwickeln wir eine ausgereifte Implementierung dieses Rahmenwerks und damit aller darin ausgedrückten Verfahren im Verifizierer CPACHECKER [BK11]. Mit Hilfe dieser Implementierung evaluieren wir zum ersten Mal systematisch 120 verschiedene Konfigurationen von Ansätzen zur automatischen Software-Verifikation und gewinnen Erkenntnisse, die für Forscher und Nutzer gleichermaßen nützlich sind. Um die Zuverlässigkeit des dafür nötigen Benchmarkings zu gewährleisten, setzen wir moderne Technologien in einem neu entwickelten Benchmarking-Tool ein.

### 1.2 Replizierbarkeit der Forschungsergebnisse

Alle in der Dissertation vorgestellten Tools sind Open Source. Um die Replizierbarkeit der Ergebnisse sicherzustellen, stehen alle Eingabedaten der Experimente, die verwendeten Tools, sowie die Ergebnisse im Rohformat zum Download bereit [3].

### 1.3 Verwandte Arbeiten

Eine Vereinheitlichung von prädikatenbasierten Verfahren zur automatischen Software-Verifikation sowie detaillierte experimentelle und konzeptionelle Vergleiche dieser Ver-

[3] `https://www.sosy-lab.org/research/phd/wendler/`

fahren existierten bisher nicht. Vorreiter für eine solche Vereinheitlichung ist das CPA-Konzept [BHT07], das ein gemeinsames Rahmenwerk für statische Analysen und Model-Checking schafft, und hier verwendet wird. Vier Verifikationsverfahren werden in Abschnitt 3 vorgestellt, ein Überblick über weitere verwandte Verfahren und Verifizierer findet sich in der Dissertation [We17, Kapitel 7].

### 1.4 Hintergrund: Model-Checking

Model-Checking ist ein gängiges Verfahren zur Verifikation von Software. Hierbei wird ein abstraktes Modell des konkreten Programms erzeugt und für dieses geprüft, ob die Spezifikation erfüllt ist. Die Konstruktion des abstrakten Modells stellt sicher, dass falls das Modell die Spezifikation erfüllt, dies auch für das konkrete Programm gilt. Eine Spezifikationsverletzung im Modell impliziert jedoch nicht notwendigerweise eine solche im Programm (Überapproximation; solche Spezifikationsverletzungen werden „unecht“ genannt). Die Prüfung des abstrakten Modells kann z. B. durch eine Erreichbarkeitsanalyse auf abstrakten Zuständen erfolgen. Ein abstrakter Zustand repräsentiert eine Menge von konkreten Zuständen des Programms. Die Nachfolgerrelation des abstrakten Modells stellt sicher, dass die Nachfolger eines abstrakten Zustands $e$ mindestens diejenigen konkreten Zustände abdecken, die das Programm ausgehend von den durch $e$ repräsentierten konkreten Zuständen in einem Schritt erreichen kann. Durch Fixpunktiteration werden alle erreichbaren abstrakten Zustände aufgezählt und auf Spezifikationsverletzungen überprüft.

Die Art des abstrakten Modells und die Tatsache, ob es endlich ist, hängt von der Wahl der zugrunde liegenden abstrakten Domäne ab, die festlegt wie abstrakte Zustände repräsentiert werden. Gängige Beispiele sind Wertedomänen (abstrakte Zustände weisen einer Teilmenge von Programmvariablen je einen konkreten Wert zu), Intervaldomänen (abstrakte Zustände weisen jeder Programmvariablen Intervalle zu) und Prädikatendomänen (abstrakte Zustände bestehen aus einer booleschen Formel über Programmvariablen).

Das Abstraktionsniveau des abstrakten Modells ist entscheidend für den Erfolg der Verifikation: ein zu abstraktes Modell kann durch die Überapproximation zu unechten Gegenbeispielen und damit falschen Ergebnissen führen, ein zu konkretes Modell kann den Aufwand stark erhöhen. Um das passende Abstraktionsniveau zu finden, kann Gegenbeispiel-gesteuerte Abstraktionsverfeinerung (CEGAR) [Cl03] genutzt werden. Dabei wird mit einem sehr abstrakten Modell begonnen und bei jeder gefundenen Spezifikationsverletzung anhand des konkreten Programms überprüft, ob diese echt oder unecht ist. Bei einem unechten Gegenbeispiel wird das abstrakte Modell so weit wie nötig verfeinert (konkretisiert). Dieser Prozess liefert entweder eine echte Spezifikationsverletzung, einen Beweis, dass keine Spezifikationsverletzung im Programm existiert, oder terminiert nicht.

## 2 Eine flexible Domäne basierend auf Prädikaten

Wir definieren eine flexible und konfigurierbare abstrakte Domäne, die Programmzustände mit Hilfe von Prädikaten über Programmvariablen repräsentiert. Dabei drücken wir die

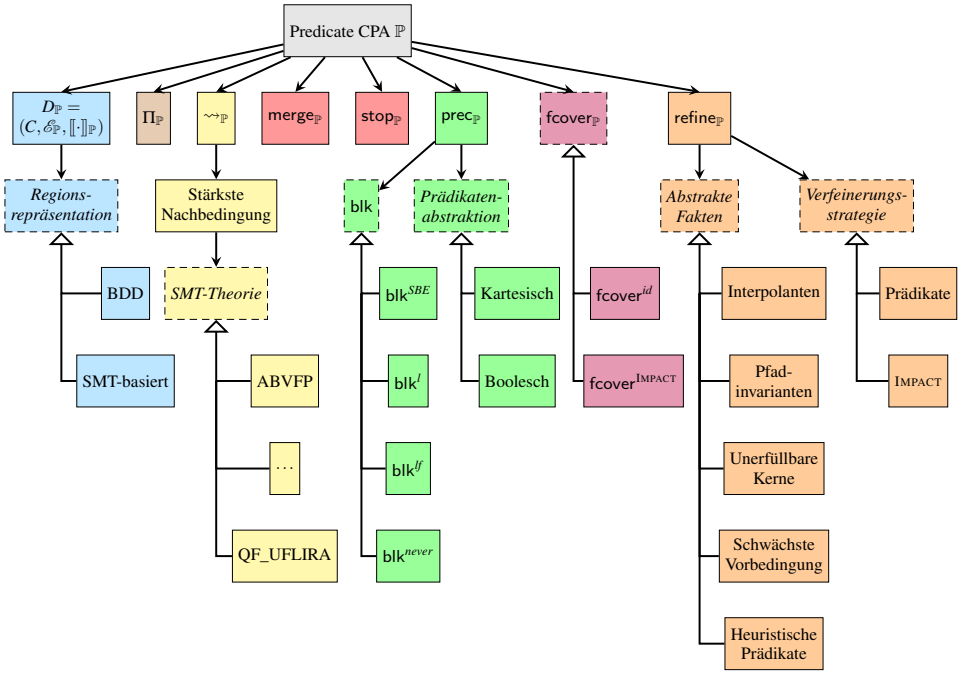



Abb. 1: Komponenten der Predicate CPA (aus [We17, BDW17])

abstrakte Domäne als konfigurierbare Programmanalyse (CPA) [BHT07] aus und nennen sie daher „*Predicate CPA*“. Die Predicate CPA besteht im Wesentlichen aus einem Halbverband $\mathscr{E}_{\mathbb{P}}$ von abstrakten Zuständen, einer Transferrelation $\rightsquigarrow_{\mathbb{P}}$ und den Operatoren $\mathsf{merge}_{\mathbb{P}}$, $\mathsf{stop}_{\mathbb{P}}$, $\mathsf{prec}_{\mathbb{P}}$, $\mathsf{fcover}_{\mathbb{P}}$ und $\mathsf{refine}_{\mathbb{P}}$. Einen Überblick über die verschiedenen Komponenten der Predicate CPA gibt Abb. 1. Um die Predicate CPA flexibel einsetzen zu können, sind bestimmte Komponenten austauschbar. In der Abbildung sind diese durch gestrichelte Boxen mit den darunterliegenden möglichen Varianten dargestellt. Im Folgenden werden beispielhaft einige wichtige Komponenten erläutert; eine vollständige Definition der Predicate CPA ist in der Dissertation [We17, Kapitel 4] enthalten.

Die abstrakten Zustände des Halbverbands $\mathscr{E}_{\mathbb{P}}$ bestehen aus booleschen Formeln über Prädikaten. Diese Formeln können entweder als binäres Entscheidungsdiagramm (BDD, aufwändig in der Erstellung, günstig in der Verwendung) oder als SMT-Formeln (günstig in der Erstellung, prüfen der Erfüllbarkeit ist teuer) repräsentiert werden. Die Transferrelation $\rightsquigarrow_{\mathbb{P}}$ ordnet einem abstrakten Zustand seine Nachfolgerzustände zu. Dies geschieht über die stärkste Nachbedingung und ohne die Berechnung einer Abstraktion. Je nach Einsatzzweck und verfügbaren Solvern können verschiedene Theorien zum Ausdrücken der stärksten Nachbedingung verwendet werden, von denen manche wie die Kombination von Feldern, Bitvektoren und Gleitkommazahlen es ermöglichen die Programmsemantik exakt abzubilden, während andere wie lineare Arithmetik nur eine Annäherung an die Programmsemantik abbilden können. Der Operator $\mathsf{prec}_{\mathbb{P}}$ kann einen abstrakten Zustand durch die Berechnung einer Prädikatenabstraktion [GS97] mit Hilfe eines SMT-Solvers

über eine gegebene Menge von Prädikaten weiter abstrahieren. Die Abstraktionsberechnung findet nur am Ende eines Blocks von Programmanweisungen statt, wobei die Blockgröße durch die Auswahl des Operators blk konfiguriert werden kann. Eine übliche Blockgröße ist z. B. die Wahl der größtmöglichen schleifenfreien Programmteile als Blöcke.

Ein weiterer wichtiger Operator ist der Operator $\mathsf{refine}_{\mathbb{P}}$, der den Verfeinerungsschritt des CEGAR-Ansatzes umsetzt. $\mathsf{refine}_{\mathbb{P}}$ prüft für jedes während der Analyse im abstrakten Modell gefundene Gegenbeispiel mit Hilfe eines SMT-Solvers, ob dieses Gegenbeispiel (d. h. ein Programmpfad zu einer Spezifikationsverletzung) im Programm valide ist. Falls dies zutrifft, ist eine Spezifikationsverletzung gefunden und die Analyse terminiert. Andernfalls ist das Gegenbeispiel unecht und die Genauigkeit der Analyse muss durch Verfeinerung des abstrakten Modells erhöht werden. Dazu berechnet $\mathsf{refine}_{\mathbb{P}}$ aus dem Gegenbeispiel passende abstrakte Fakten, die dem abstrakten Modell hinzugefügt werden müssen. Hierfür kann z. B. Craig-Interpolation [Cr57] genutzt werden. Anschließend wird das abstrakte Modell verfeinert, z. B. durch das Hinzufügen von Prädikaten aus den abstrakten Fakten zu jener Menge, die für die Abstraktionsberechnungen verwendet wird, oder durch das Stärken von abstrakten Zuständen durch Hinzukonjugieren der passenden abstrakten Fakten.

## 3 Vereinheitlichung von prädikatenbasierten Algorithmen

Aufbauend auf der Predicate CPA können wir nun ein Rahmenwerk definieren, das mehrere vorhandene prädikatenbasierte Algorithmen vereinheitlicht. Durch das Ausdrücken als CPA können wir den CPA-Algorithmus zur abstrakten Erreichbarkeitsanalyse [BHT07] mit nur wenigen Modifikationen verwenden [We17, BDW17]. Dieser Algorithmus berechnet die Menge aller erreichbaren abstrakten Zustände einer abstrakten Domäne wie der Predicate CPA durch eine Fixpunktiteration über die Transferrelation. Wir verwenden dabei die Predicate CPA in einer an das jeweils gewünschte Verfahren angepassten Konfiguration, d. h. wir wählen für jede der austauschbaren Komponenten (in Abb. 1 gestrichelt) eine bestimmte Variante aus. Außerdem setzen wir weitere auf dem CPA-Algorithmus aufbauende Algorithmen ein, die in der Dissertation detailliert beschrieben [We17, Kapitel 5] sind.

### 3.1 Bounded Model-Checking

Für Bounded Model-Checking [Bi99], d. h. das $k$-malige Abrollen des Programms und das Prüfen der Spezifikation für alle so erhaltenen endlichen Pfade, wird eine Konfiguration der Predicate CPA gewählt, die $\mathsf{blk}^{never}$ nutzt und ansonsten beliebig ist. Durch $\mathsf{blk}^{never}$ ist die Blockgröße unendlich groß und das gesamte Programm besteht aus einem einzigen Block. Darüber hinaus muss durch Kombination der Predicate CPA mit einer weiteren CPA, die die Beschränkung definiert (z. B. maximal $k$ Schleifenabrollungen), sichergestellt werden, dass der CPA-Algorithmus terminiert. Anschließend wird durch einen einmaligen Aufruf eines SMT-Solvers überprüft, ob auf einem der gefundenen Pfade eine Spezifikationsverletzung erreichbar ist. Die dazu nötige Formel kann aus den abstrakten Zuständen abgelesen werden, die durch die unendliche Blockgröße nie abstrahiert wurden und daher die Programmsemantik (für den analysierten Teil) exakt abbilden.

### 3.2 $k$-Induktion

Mit Bounded Model-Checking können nur Spezifikationsverletzungen gefunden werden, jedoch im Allgemeinen keine Erfüllung der Spezifikation bewiesen werden. Dies lässt sich durch einen $k$-Induktionsbeweis [DKR11] durchführen. Hierbei besteht der Induktionsanfang aus der Überprüfung, dass in den ersten $k$ Schleifeniterationen die Spezifikation erfüllt ist. Dies ist identisch zu Bounded Model-Checking mit Tiefe $k$. Für den Induktionsschritt wird überprüft, ob in einer beliebigen Schleifeniteration eine Spezifikationsverletzung existieren kann, falls in den vorhergehenden $k$ Iterationen keine Spezifikationsverletzung aufgetreten ist (Induktionsannahme). Dies lässt sich umsetzen durch Bounded Model-Checking mit Tiefe $k+1$ und einem initialen Zustand am Kopf der Schleife des Programms anstatt am Programmanfang (das Verfahren lässt sich auf Programme mit mehreren Schleifen erweitern). Daher lässt sich auch dieses Verfahren mit Hilfe der Predicate CPA durchführen.

### 3.3 Prädikatenabstraktion

Späte Prädikatenabstraktion [GS97, He02] ist ein klassisches Verfahren für Model-Checking und wird durch die Predicate CPA unterstützt, indem für $\mathsf{blk}$ eine Wahl außer $\mathsf{blk}^{never}$ getroffen wird und die Prädikate-Verfeinerungsstrategie genutzt wird. Außerdem werden typischerweise die abstrakten Zustände durch BDDs repräsentiert, ansonsten ist die Konfiguration beliebig. Damit der $\mathsf{refine}_{\mathbb{P}}$-Operator zum Zuge kommt, muss ein Algorithmus für Gegenbeispiel-basierte Abstraktionsverfeinerung zusätzlich zum CPA-Algorithmus verwendet werden. Hierbei ist die Menge der Prädikate, die von $\mathsf{prec}_{\mathbb{P}}$ zur Abstraktionsberechnung verwendet werden, am Anfang leer, so dass jede Abstraktionsberechnung maximal überapproximiert. Bei unechten Gegenbeispielen werden dann durch $\mathsf{refine}_{\mathbb{P}}$ passende Prädikate gefunden und in die Menge der Prädikate eingefügt, wodurch zukünftige Abstraktionsberechnungen genauer werden und die unechten Gegenbeispiele ausgeschlossen werden können. Außerdem werden nach der Verfeinerung die zu ungenauen Teile des abstrakten Modells verworfen und neu berechnet.

### 3.4 IMPACT

Der IMPACT-Algorithmus [Mc06] ist ein Verfahren zur Gegenbeispiel-gesteuerten Abstraktionsverfeinerung, der gegenüber der Prädikatenabstraktion effizienter arbeiten soll, indem er die teuren Abstraktionsberechnungen vermeidet. Obwohl in der initialen Beschreibung komplett anders dargestellt, lässt er sich in unserem Rahmenwerk sehr ähnlich zur Prädikatenabstraktion ausdrücken: Die wichtigste Abweichung ist die IMPACT-Verfeinerungsstrategie, die das abstrakte Modell verfeinert, indem es die gefundenen abstrakten Fakten zu den vorhandenen abstrakten Zuständen hinzukonjugiert. Dadurch ist kein Verwerfen und Neuberechnen des abstrakten Modells nötig, und da die Menge der Prädikate für die Abstraktionsberechnung dauerhaft leer bleibt, bleibt die Abstraktionsberechnung dauerhaft eine triviale Operation. Allerdings werden nun statt BDDs SMT-Formeln zur Repräsentation der abstrakten Zustände verwendet, was dazu führt, dass Operationen wie der Vergleich von abstrakten Zuständen (für die Fixpunkterkennung nötig) teurer werden.

### 3.5 Anwendungen

Unser Rahmenwerk für prädikatenbasierte Software-Verifikation, das auf der Predicate CPA basiert, erlaubt es also so unterschiedliche Verfahren wie die genannten nur durch wenige Konfigurationsänderungen als Varianten auszudrücken. Die Unterschiede und Gemeinsamkeiten der Verfahren werden so hervorgehoben. Dies erlaubt es uns neue Erkenntnisse über die Kernideen dieser Verfahren zu erlangen [BW12]. Außerdem werden dadurch experimentelle Vergleichsstudien in einer neuen Qualität möglich [BDW17], da alle irrelevanten Unterschiede eliminiert wurden. Darüber hinaus können wir nun ohne weiteren Aufwand neue Verfahren definieren und damit experimentieren, indem wir die vorhandenen Verfahren und ihre Bestandteile neu kombinieren. Solche Ansätze werden in der Dissertation beschrieben [We17, Abschnitt 5.4].

## 4 Benchmarking für zuverlässige experimentelle Forschung

Die experimentelle Evaluation der Effizienz durch Benchmarking, d. h. das Messen von Zeit- und Speicherverbrauch eines Tools, ist eine effektive und günstige Methode um Forschungsergebnisse zu beurteilen. Im Bereich der Software-Verifikation ist dies die Standardmethode, z. B. um verschiedene Algorithmen oder Tools anhand ihrer Effizienz für eine große Menge an zu verifizierenden Programmen zu vergleichen. Im Rahmen der Dissertation wollen wir die in Abschnitt 3 vorgestellten Verfahren evaluieren und unsere Implementierung mit anderen Verifizierern vergleichen. Hierzu benötigen wir die Möglichkeit, den Zeit- und Speicherverbrauch einer Ausführung eines beliebigen Programms exakt zu messen. Um die Replizierbarkeit der Ergebnisse sicherzustellen, ist es darüber hinaus nötig, während der Programmausführung den Speicherverbrauch auf ein festgesetztes Maß zu limitieren und das Programm gegenüber äußeren Einflüssen abzuschirmen. Inbesondere Effekte, die zu einer nicht-deterministischen Zeitmessung führen können, müssen soweit wie möglich ausgeschlossen werden.

Drei Hauptprobleme müssen wir für zuverlässiges Benchmarking berücksichtigen. Erstens starten viele Programme Kindprozesse zur Erledigung bestimmter Aufgaben, deren Ressourcenverbrauch mit berücksichtigt werden muss. Kindprozesse zu nutzen ist eine gängige Praxis von Verifizierern, aber gängige Standardverfahren zum Messen des Zeitverbrauchs von Prozessen messen die Zeiten von Kindprozessen nicht zuverlässig mit. Dies kann zu Messfehlern beliebiger Größe führen. Zweitens kommt es zu Problemen, wenn aus Zeitgründen das Benchmarking mehrerer unabhängiger Programmausführungen parallel auf der gleichen Maschine nötig ist. Hardwareeigenschaften gängiger Systeme wie Hyper-Threading, Turbo Boost und Nonuniform Memory Access (NUMA) können dazu führen, dass der Zeitverbrauch eines Programms von den Aktivitäten parallel laufender Prozesse abhängt, selbst wenn jeder Prozess spezifische Hardwareeinheiten (z. B. CPU-Kerne) exklusiv zugewiesen bekommen hat. Daher muss beim Benchmarking darauf geachtet werden, die Zuteilung der Hardwareressourcen auf die Prozesse so vorzunehmen, dass diese Einflüsse ausgeschlossen oder zumindest minimiert werden. Die dritte Art von Problemen besteht darin, die Unabhängigkeit mehrerer (sequentieller oder paralleler) Programmausführungen zu gewährleisten, auch wenn das Programm z. B. temporäre Dateien

an bestimmte feste Orte schreibt. Insgesamt können wir die sechs in Abb. 2 aufgeführten Anforderungen identifizieren, die diese Probleme adressieren und die von einem Benchmarking-System sichergestellt werden müssen.

1. Messe und limitiere den Ressourcenverbrauch akkurat
2. Beende Prozesse zuverlässig
3. Weise CPU-Kerne gezielt zu
4. Respektiere Nonuniform Memory Access
5. Vermeide die Nutzung von Auslagerungsspeicher
6. Isoliere individuelle Programmausführungen

Abb. 2: Anforderungen für zuverlässiges Benchmarking

Diese Anforderungen können unter Linux nur umgesetzt werden durch die Verwendung von Kontrollgruppen („control groups“ oder „cgroups“) und Namensräumen („namespaces“), zwei Features des Linux-Kernels. Mit Kontrollgruppen lassen sich Prozesse in Gruppen zusammenfassen, und Ressourcen wie Zeit und Speicher lassen sich für solche Gruppen messen und limitieren. Mit Namensräumen lassen sich Prozesse in sogenannten „Containern“ isolieren, z. B. lässt sich verhindern, dass bestimmte Prozesse mit anderen Prozessen oder mit dem Netzwerk kommunizieren. Darüber hinaus lässt sich für Container die Sicht auf das Dateisystem modifizieren, so dass z. B. bestimmte Verzeichnisse unsichtbar sind oder jeder Container sein eigenes Verzeichnis für temporäre Dateien bekommt, um Wechselwirkungen zwischen Containern auszuschließen. Container sind ähnlich zu virtuellen Maschinen, allerdings ohne die Geschwindigkeitseinbußen und den höheren Speicherverbrauch von virtuellen Maschinen.

Da gängige Benchmarking-Programme die Anforderungen für zuverlässiges Benchmarking nicht einhalten, stellen wir das Tool BENCHEXEC [We17, BLW17] zur Verfügung. Es benutzt Kontrollgruppen und Namensräume, um die Einhaltung der Anforderungen aus Abb. 2 umzusetzen. BENCHEXEC ist unter der Open-Source-Lizenz Apache 2.0 auf GitHub verfügbar[4]. Es ist die technische Grundlage der International Competition on Software Verification (SV-COMP) [Be16], bei der über mehrere Jahre Millionen von Ausführungen mehrerer Dutzend Verifizierer durchgeführt und gemessen wurden. Dabei wäre es ohne die Techniken, auf denen BENCHEXEC basiert, stellenweise zu großen Messfehlern gekommen. Dies zeigt, dass zuverlässiges Benchmarking unersetzlich ist, um falsche wissenschaftliche Schlüsse zu vermeiden. Unser Tool BENCHEXEC ermöglicht dies.

## 5 Experimentelle Evaluation

Eine experimentelle Studie, die die vier in Abschnitt 3 vorgestellten Verfahren mit Hilfe unseres Rahmenwerks untereinander vergleicht, wurde bereits separat durchgeführt [BDW17], und lieferte interessante Erkenntnisse über die Effizienz und Effektivität der Verfahren für bestimmte Klassen von Eingabeprogrammen. In der Dissertation wird daher zum einen evaluiert, wie die erstellte Implementierung des Rahmenwerks in CPACHECKER im Vergleich

[4] `https://github.com/sosy-lab/benchexec`

mit anderen Verifizierern (basierend auf den gleichen und anderen Verfahren) abschneidet. Zum Vergleich werden die Ergebnisse der Int. Competition on Software Verification 2017 (SV-COMP'17) herangezogen, dem größten internationalen Wettbewerb für Software-Verifizierer mit 32 Teilnehmern im Jahr 2017. Der Vergleich wird für 5 594 C-Programme (mit und ohne bekannten Spezifikationsverletzungen) mit insgesamt 73 Millionen Zeilen Code (2.4 GB) ausgeführt. Wir zeigen, dass die Implementierung auf höchstem Niveau liegt, indem die verschiedenen Varianten der Predicate CPA jeweils vergleichbar viele Programme korrekt verifizieren wie die besten Teilnehmer an SV-COMP'17. In der Zwischenzeit wird dies bestätigt durch eine erfolgreiche Teilnahme an SV-COMP'18, in der CPACHECKER mit Hilfe dieser Implementierung den Sieg in der Kategorie Overall erzielte.

Zum anderen wurde mit den gleichen Eingabeprogrammen eine Studie mit 120 verschiedenen Konfigurationen der Predicate CPA durchgeführt – 30 Untervarianten für jedes der vier vorgestellten Verfahren. Diese Studie umfasste insgesamt 671 280 Ausführungen des Verifizierers und benötigte 3 620 Tage Rechenzeit. Die Studie zeigt, dass auch einige technische und kleinere konzeptionelle Entscheidungen, die in der Theorie unabhängig vom verwendeten Verifikationsverfahren sind, nicht nur signifikanten Einfluss auf die Effektivität und Effizienz haben, sondern auch Vergleiche zwischen den Verfahren beeinflussen können, da manche dieser Entscheidungen übermäßig nachteilig bei bestimmten Verfahren wirken. Die Ergebnisse der Studie liefern daher nicht nur wertvolle Informationen für Nutzer, die effektive und effiziente Verifikationsverfahren benötigen, sondern auch entscheidende Hinweise für Forscher, die diese bei experimentellen Studien beachten müssen. Ohne die Vereinheitlichung der Verifikationsverfahren in einem gemeinsamen flexiblen Rahmenwerk wäre diese Studie nicht möglich gewesen.

## 6 Zusammenfassung

Das in der Dissertation vorgestellte Rahmenwerk vereinheitlicht zum ersten Mal verschiedenste Verfahren zur automatischen Software-Verifikation und löst das Problem des fragmentierten Stands der Forschung, unterstützt durch eine robuste und effiziente Implementierung auf weltweit höchstem Niveau. Dies liefert fruchtbare Erkenntnisse und bildet die Grundlage für neue konzeptionelle und experimentelle Möglichkeiten sowohl in der Forschung als auch im praktischen Einsatz von prädikatenbasierten Analysen. Das vorgestellte neuartige Benchmarking-System ermöglicht mit Hilfe moderner Technologien zuverlässige experimentelle Forschung.

Sowohl das präsentierte Rahmenwerk als auch die entwickelten Tools haben sich bereits in einer Reihe von darauf aufbauenden Forschungsprojekten verschiedener Gruppen als hilfreich erwiesen. Die Verfahren und Tools der Dissertation werden regelmäßig zur Verifikation von Gerätetreibern des Linux-Kernels eingesetzt und haben dabei bereits geholfen Dutzende von Fehlern zu finden. Die in dem weltweit wichtigsten Wettbewerb für Software-Verifizierer erzielten Erfolge wurden mit der Kurt-Gödel-Medaille ausgezeichnet.

## Literaturverzeichnis

[BDW17] Beyer, D.; Dangl, M.; Wendler, P.: A Unifying View on SMT-Based Software Verification. J. Autom. Reasoning, 2017.

[Be16] Beyer, D.: Reliable and Reproducible Competition Results with BENCHEXEC and Witnesses (SV-COMP 2016). In: Proc. TACAS. LNCS 9636. Springer, S. 887–904, 2016.

[BHT07] Beyer, D.; Henzinger, T. A.; Théoduloz, G.: Configurable Software Verification: Concretizing the Convergence of Model Checking and Program Analysis. In: Proc. CAV. LNCS 4590. Springer, S. 504–518, 2007.

[Bi99] Biere, A.; Cimatti, A.; Clarke, E. M.; Zhu, Y.: Symbolic Model Checking without BDDs. In: Proc. TACAS. LNCS 1579. Springer, S. 193–207, 1999.

[BK11] Beyer, D.; Keremoglu, M. E.: CPACHECKER: A Tool for Configurable Software Verification. In: Proc. CAV. LNCS 6806. Springer, S. 184–190, 2011.

[BLW17] Beyer, D.; Löwe, S.; Wendler, P.: Reliable Benchmarking: Requirements and Solutions. Int. J. Softw. Tools Technol. Transfer, 2017.

[BW12] Beyer, D.; Wendler, P.: Algorithms for Software Model Checking: Predicate Abstraction vs. IMPACT. In: Proc. FMCAD. FMCAD, S. 106–113, 2012.

[Cl03] Clarke, E. M.; Grumberg, O.; Jha, S.; Lu, Y.; Veith, H.: Counterexample-guided abstraction refinement for symbolic model checking. J. ACM, 50(5):752–794, 2003.

[Cr57] Craig, W.: Linear Reasoning. A New Form of the Herbrand-Gentzen Theorem. J. Symb. Log., 22(3):250–268, 1957.

[DKR11] Donaldson, Alastair F.; Kroening, Daniel; Rümmer, Philipp: Automatic analysis of DMA races using model checking and *k*-induction. FMSD, 39(1):83–113, 2011.

[GS97] Graf, S.; Saïdi, H.: Construction of Abstract State Graphs with PVS. In: Proc. CAV. LNCS 1254. Springer, S. 72–83, 1997.

[He02] Henzinger, T. A.; Jhala, R.; Majumdar, R.; Sutre, G.: Lazy abstraction. In: Proc. POPL. ACM, S. 58–70, 2002.

[Mc06] McMillan, K. L.: Lazy Abstraction with Interpolants. In: Proc. CAV. LNCS 4144. Springer, S. 123–136, 2006.

[We17] Wendler, Philipp: Towards Practical Predicate Analysis. Dissertation, Univ. Passau, 2017.

**Philipp Wendler** wurde am 29. April 1985 in Roth geboren. Er studierte von 2005 bis 2010 Informatik an der Universität Passau und erlangte in dieser Zeit sowohl den Bachelor- als auch den Master-Abschluss, jeweils mit Auszeichnung. Im Rahmen seiner Master-Arbeit, die mit dem IHK-Preis der IHK Niederbayern ausgezeichnet wurde, beschäftigte er sich erstmals mit Software-Verifikation. Anschließend vertiefte er die Forschung in diesem Bereich im Rahmen einer Promotion am Lehrstuhl von Prof. Dr. Dirk Beyer an der Universität Passau. Insbesondere beschäftigte er sich mit Software Model Checking, wo er ein Rahmenwerk für die Vereinheitlichung einer breiten Reihe von Ansätzen entwickelte. Er ist einer der Hauptentwickler des CPACHECKER-Projekts, eines der erfolgreichsten Software-Verifizierer weltweit. Für seine Arbeiten wurde er mit dem Young Scientist Award und der Kurt-Gödel-Medaille ausgezeichnet. Die Promotion schloss er 2017 mit Auszeichnung ab und seit 2016 arbeitet er bei Prof. Dr. Dirk Beyer an der LMU München.

# Eine Datenspezifikationsarchitektur

Stefan Widmann[1]

**Abstract:** Eingebettete Systeme werden in zunehmendem Maße in sicherheitsgerichteten Anwendungen eingesetzt, so z. B. in Form des Lenkens „steer-by-wire“ und Bremsens „brake-by-wire“ im Automobilbereich. Die Komplexität der in diesen Systemen eingesetzten Hard- und Software steigt, und die fortwährende Verkleinerung der Strukturbreiten in integrierten Schaltkreisen macht diese immer empfindlicher gegenüber Umgebungseinflüssen wie Strahlung, was sich in einer steigenden Wahrscheinlichkeit von Datenflussfehlern auswirkt. Anstatt dem Trend zur Fehlererkennung durch immer komplexere Software auf konventionellen Prozessorarchitekturen zu folgen, wurde in der Arbeit eine neuartige Prozessorarchitektur mit umfassenden hardwarebasierten Fehlererkennungsmerkmalen vorgestellt. Diese ermöglichen die einfache und zuverlässige Erkennung von zur Laufzeit auftretenden Datenflussfehlern, wodurch die Architektur bisherigen Ansätzen deutlich überlegen ist.

## 1 Einführung

Es besteht deutlicher Bedarf an der Verbesserung heutiger Fehlervermeidungs- und -erkennungsmaßnahmen, die in sicherheitsgerichteten Systemen zum Einsatz kommen, also Systemen, die Verantwortung für Mensch, Umwelt und Investitionsgüter tragen. Es ist notwendig, Fehler weitestgehend zu vermeiden, und trotz aller Vermeidungsmaßnahmen trotzdem auftretende Fehler so frühzeitig wie möglich zu erkennen: idealerweise im Moment ihres Auftretens, und nicht erst, wenn ihre Auswirkungen (z. B. durch Vergleich von Ergebnissen oder Zwischenergebnissen) sichtbar werden und ggf. bereits zu gefährlichen Ausgaben bis hin zu fatalen Unfällen führen. Die möglichen Folgen nicht erkannter oder nicht korrekt behandelter Fehler, die während Spezifikation, Entwurf oder Implementierung von Hard- oder Software in ein System eingebracht wurden oder zur Laufzeit auftreten, zeigen einige Vorfälle aus der Vergangenheit eindrücklich:

Recht bekannt ist die Selbstzerstörung der Rakete Ariane 5, die hohen Sachschaden verursachte. Nach [Li96] kam es bei der Umwandlung einer 64-Bit-Gleitkommazahl in eine vorzeichenbehaftete 16-Bit-Ganzzahl zu einer Überschreitung des Wertebereichs, weil die horizontale Geschwindigkeit der Ariane 5 deutlich höher war als die der Ariane 4. Der Fehler wurde als solcher erkannt und ein Diagnose-Bitmuster an den Bordrechner gesendet, der dieses fehlerhafterweise als Flugdaten interpretierte und die Ablenkdüsen voll aussteuerte, was in der Folge die Selbstzerstörung der Rakete auslöste.

Ebenfalls großen Sachschaden verursachte der Verlust des Mars Climate Orbiter MCO, einer Sonde der NASA. Nach [St99] war die Trägerrakete durch die Verwendung inkompatibler Einheiten – eine Softwarekomponente rechnete in metrischen SI-Einheiten, die andere in angloamerikanischen Maßeinheiten – zum Zeitpunkt des Absetzens der Sonde im Marsorbit rund 170 km zu tief, was entweder zu deren Abdriften oder Verglühen führte.

[1] FernUniversität in Hagen, stefan.widmann@gmx.de

Beim Einsatz des für medizinische Zwecke genutzten Teilchenbeschleunigers Therac-25 kam es zu mindestens sechs schweren Unfällen, bei denen nach [LT93] Patienten aufgrund von Softwarefehlern massiven Strahlungsüberdosen mit teils tödlichen Folgen ausgesetzt wurden. Während das Vorgängergerät Therac-20 noch zusätzliche hardwaretechnische Sicherheitseinrichtungen und mechanische Verriegelungen nutzte, wurden diese Maßnahmen beim Therac-25 durch eine reine Softwarelösung ersetzt – so groß war das Vertrauen in die Software des Geräts. Die Unfälle wurden dadurch ausgelöst, dass es bei entsprechender Nutzereingabe aufgrund unzureichender Synchronisierungsmechanismen zu einer Vermischung vorhergehender und aktueller Betriebsparameter kommen konnte, wodurch sich unzulässige Kombinationen von Bestrahlungsart, -dosen und -zeiten ergaben.

Konventionelle Prozessorarchitekturen sind vor allem auf maximalen Datendurchsatz hin ausgelegt, nicht auf Einfachheit, Fehlervermeidung und -erkennung. Diese – eigentlich ungeeigneten Systeme – kommen meist aus ökonomischen Gründen in den beschriebenen Anwendungen zum Einsatz und werden als „commercial-off-the-shelf (COTS)“ bezeichnet. Die Datenworte in den Speichern dieser Systeme enthalten neben dem eigentlichen Datenwert keinerlei weiterführende Informationen, die dessen Eigenschaften beschreiben. In solchen Architekturen und Systemen kann den steigenden Anforderungen und Fehlerwahrscheinlichkeiten nur durch weiter steigende Komplexität der Software begegnet werden, z. B. durch den Einsatz von softwarebasierter arithmetischer Kodierung wie Software Encoded Processing (SEP) und Compiler Encoded Processing (CEP) [Sc11].

In [Go14] wurde der Frage nachgegangen, wie sich der Kontrollfluss innerhalb sicherheitsgerichteter Echtzeitsysteme mit möglichst einfachen hardwarebasierten Mitteln überwachen lässt. Allerdings wird in [Te87] davon gesprochen, dass 80 - 90 % aller Programmfehler Datenflussfehler sind und nur die verbleibenden 10 - 20 % auf Kontrollflussfehler entfallen.

In der diesem Beitrag zugrundeliegenden Arbeit [Wi17] wurden 20 datenflussbezogene Fehler- und Angriffsarten identifiziert und eine neuartige, leistungsfähige Prozessorarchitektur für die hardwarebasierte Überwachung des Datenflusses innerhalb von sicherheitsgerichteten Echtzeitsystemen vorgestellt. Sie fügt Datenwerten eine umfassende, hardwareles- und -prüfbare Beschreibung ihrer Eigenschaften hinzu, die untrennbar mit den Datenwerten verbunden ist und mit ihnen gespeichert und verarbeitet wird. Dadurch ist die Architektur in der Lage, alle 20 Fehler- und Angriffsarten zur Laufzeit aufzudecken und entsprechende Fehlerbehandlungmaßnahmen einzuleiten und ist bisherigen Ansätzen damit deutlich überlegen.

## 2 Der Stand von Wissenschaft und Technik und dessen Nachteile

Die konventionellen Architekturen x86 im geschützten und 64-Bit-Modus [In91, AM12] und ARM [AR11] können trotz größter Bemühungen, unter Nutzung komplexester Maßnahmen, einen maximalen Datendurchsatz zu bieten, nur wenige datenflussbezogene Fehler- und Angriffsarten erkennen.

Die auf den konventionellen Architekturen basierenden, für sicherheitsgerichtete Systeme spezialisierten Prozessoren, wie z. B. der TI Hercules [Te15], sind in der Lage, mehr Fehlerarten aufzudecken, basierend auf dem Einsatz gewisser Redundanz- und Diversitätsarten. Allerdings ist die Reichweite der Fehlererkennung begrenzt und viele Fehlerarten bleiben weiterhin unerkannt.

Die weitestgehend in Vergessenheit geratenen Architekturarten Datentyp-, Datenstruktur- und Befähigungsarchitekturen [Gi93, Fe73, AE] fügen Speicherinhalten Kennungen hinzu, die hardwareverständlich die Inhalte des Speichers beschreiben. Damit können durch die Architekturen bestimmte Fehlerarten erkannt werden, doch auch hier bleiben viele weitere unentdeckt.

Datenflussarchitekturen [Gi93] sind auf die Bearbeitung von Datenflüssen spezialisiert, wodurch bestimmte Fehlerarten in der Theorie grundsätzlich nicht auftreten können. Werden diese Architekturen jedoch auf unterster Ebene wiederum durch konventionelle Architekturen realisiert, so z. B. die Verarbeitungseinheiten in [Ł95], dann besteht die genannte Beschränkung der Fehlerbandbreite nicht mehr. Bei der Verarbeitung der Daten findet keinerlei Prüfung derer Eigenschaften statt, wodurch Fehlerarten wie z. B. inkompatible Datentypen der Operanden nicht erkannt werden können.

Das Fehlererkennungsmerkmal Application Data Integrity ADI, später Silicon Secured Memory SSM genannt, des Oracle SPARC M7 Prozessors [Or14] fügt 64-Byte-Datenblöcken eine Versionskennung hinzu und nutzt Teile von Zeigern, um dort eine erwartete Version zu hinterlegen. Dadurch können Abweichungen zwischen erwarteter und vorgefundener Datenversion aufgedeckt werden. Nachteilig ist, dass Versionen nur für Datenblöcke, nicht jedoch für einzelne Datenwerte vergeben werden können. Weiterhin müssen die Versionsangaben aufwendig durch die Software aktualisiert werden.

Die dynamische Datenflussprüfung DDFV [MS07] erlaubt die signaturbasierte Prüfung der Datenflüsse auf Registerebene, ist aber nicht in der Lage, in diese Prüfungen den oder die Speicher einzubeziehen, geschweige denn systemweite Datenflüsse zu überwachen. Zudem kann eine Abweichung vom vorgesehenen Datenfluss erst am Ende eines Überwachungsblocks, und nicht im Moment des Auftretens erkannt werden. Dies kann zu spät sein, und die ungültigen Ergebnisse können unter Umständen bereits zu gefährlichen Ausgaben geführt haben.

Die arithmetische AN-Kodierung [Br60] kann zusammen mit den Erweiterungen zur ANBD-Kodierung [Fo89] einige wichtige Datenflussfehler erkennen, ist jedoch nur für bestimmte Operationsarten und Datentypen geeignet und verursacht erhöhten Laufzeitbedarf. Zudem sind die kodierten Werte nicht mehr ohne Weiteres menschenlesbar, wodurch die Fehlersuche erschwert wird.

Die Datenflussüberwachung in Netzwerkprotokollen wie z. B. TCP [RF81] weist nur wenige Fehlererkennungsmerkmale auf, während Protokolle für sicherheitsgerichtete Feldbusse [IE16] deutlich mehr Fehlererkennungsmechanismen einsetzen. Allerdings wird nur die erfolgreiche Übertragung der Daten über die verschiedenen Kommunikationsstrecken geprüft, nicht jedoch deren Weiterverarbeitung innerhalb der Datenverarbeitungseinheiten.

Allgemein fehlt eine systemweite, ganzheitliche Betrachtung von Daten, ihrer Eigenschaften und ihrer Wege durch ein System. Selbst wenn bestimmte Dateneigenschaften von der Software lokal zur Laufzeit betrachtet oder sogar überwacht werden, so werden diese getrennt von den eigentlichen Daten verwaltet. Es besteht die Gefahr von Inkonsistenzen, und die Information geht bei der Übertragung der Daten zwischen verschiedenen Softwareprogrammen, spätestens jedoch bei der Übertragung der Daten an andere Systemkomponenten verloren. Die Hauptverantwortung für die Fehlererkennung trägt meist die Software, was deren Komplexität und Fehlerwahrscheinlichkeit weiter erhöht.

## 3 Ziel und Ergebnisse der Arbeit

Ein Ziel der Arbeit [Wi17] war die Identifikation relevanter datenflussbezogener Fehler- und Angriffsarten und der Eigenschaften von Daten im Anwendungsbereich sicherheitsgerichteter Echtzeitsysteme, um

- einfache Fehlervermeidungs- und -erkennungsmöglichkeiten auf Hardwareebene,
- einfache Überwachungsmöglichkeiten von Echtzeitbedingungen durch die Hardware selbst und
- eine ganzheitliche Betrachtung eines Gesamtsystems inklusive aller eingesetzten Hard- und Software

zu erreichen, ohne dem Trend zur unnötigen weiteren Erhöhung der Komplexität der eingesetzten Entwicklungswerkzeuge (z. B. Übersetzer) oder der entstehenden Software zu folgen. In Datentyp-, Datenstruktur- und Befähigungsarchitekturen wurden sehr leistungsfähige Fehlererkennungsmerkmale geboten, und dies mit einfachsten Mitteln. Es galt, diese Merkmale zu nutzen und deutlich zu erweitern.

Feustel zitiert Iliffe in [Fe72] dahingehend, dass die Eigenschaften von Datenfeldern untrennbar von den eigentlichen Daten in den Feldern selbst unterzubringen seien, anstatt in den auf die Felder zugreifenden Algorithmen. Bezog sich diese Anforderung von Iliffe zunächst nur auf die Anzahl der Elemente innerhalb des Felds und deren Datentypen, so wurde sie für die Arbeit zum folgenden zentralen Entwurfsparadigma erweitert:

> *Alle ein Datenspeicherelement beschreibenden Eigenschaften sollen untrennbar mit diesem verknüpft, gespeichert, übertragen, verarbeitet und in einer hardwareverständlichen und -überprüfbaren Form dargestellt werden.*

Im Zuge der Arbeit [Wi17] entstand – basierend auf dieser Forderung – eine neuartige Prozessorarchitektur, die aufgrund der umfassenden, hardwareverständlichen Beschreibung von Dateneigenschaften als „Datenspezifikationsarchitektur“ bezeichnet wird. Die Beiträge der Arbeit zum Stand von Wissenschaft und Technik sind dabei:

- die Identifikation von insgesamt 20 datenflussbezogenen Fehler- und Angriffsarten, die in diesem Beitrag in Tabelle 1 aufgelistet werden,
- eine umfassende Sammlung der Eigenschaften von Daten in sicherheitsgerichteten Echtzeitsystemen und
- die Vorstellung der auf dieser Grundlage entwickelten Datenspezifikationsarchitektur, welche die identifizierten Dateneigenschaften in Form von hardwareverständlichen Kennungen den Datenwerten hinzufügt.

Die Neuheiten der Fehlererkennungsmerkmale der Datenspezifikationsarchitektur gegenüber dem Stand von Wissenschaft und Technik, von denen 8 zum Patent beim Deutschen Marken- und Patentamt angemeldet wurden, sind:

- die Definition von Messwertdatentypen in Form eines Werteintervalls zur Darstellung fehlerbehafteter Werte (Abb. 1), um die Fehlerfortpflanzung bei der Werteverarbeitung durch Intervallarithmetik zu verfolgen und eventuelle Genauigkeitsprobleme zu erkennen, zusammen mit speziellen Befehlen zur Prüfung der Genauigkeit,

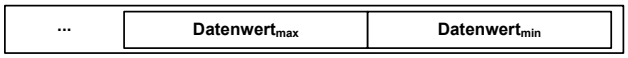



Abb. 1: Datenwert als Werteintervall

- eine Wertebereichskennung [WH17h] (Abb. 2), die es der Hardware erlaubt, einerseits eine Plausibilitätsprüfung beim Lesen eines Datenspeicherelements durchzuführen, indem geprüft wird, ob der Datenwert innerhalb des spezifizierten Wertebereichs liegt, und es ihr andererseits beim Schreiben eines Datenwerts ermöglicht, sicherzustellen, dass dieser innerhalb des spezifizierten Wertebereichs liegt,

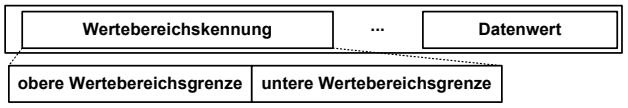



Abb. 2: Wertebereichskennung im Datenspeicherelement

- die Erweiterung der von Datentyparchitekturen bekannten Datentypkennungen um abgeleitete Unterdatentypen mit Einschränkungen der gestatteten Operationen in einer Typberechtigungskennung [WH17a] (Abb. 3), wodurch die versuchte Anwendung nicht zulässiger Operationen von der Hardware als Fehler erkannt wird,

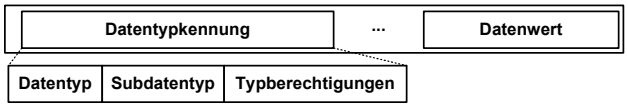



Abb. 3: Erweiterte Datentypkennung im Datenspeicherelement

- eine Einheitenkennung [WH17c] (Abb. 4), die die Einheit des Datenwerts eines Datenspeicherelements in Form von Potenzen der sieben SI-Basiseinheiten beschreibt und der Hardware umfassende Kompatibilitätsprüfungen der Operanden und die Ermittlung der Einheit des Ergebnisses einer Operation gestattet,

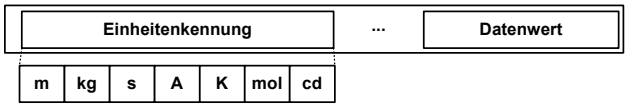



Abb. 4: Einheitenkennung im Datenspeicherelement

- eine Verarbeitungswegkennung [WH17e] (Abb. 5), die beschreibt, wer die Daten erzeugt hat, welche Stationen die Daten auf ihrem Weg von Datenquelle bis -senke verarbeiten dürfen und wer die Daten schlussendlich entgegennehmen darf, so dass die Hardware basierend auf diesen Informationen fehlgeleitete Daten zuverlässig als solche erkennen kann,

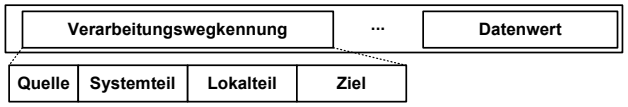



Abb. 5: Verarbeitungswegkennung im Datenspeicherelement

- eine Zeitschrittkennung [WH17f] (Abb. 6), die beschreibt, zu welchem diskreten Zeitpunkt der betroffene Datenwert generiert worden ist, zusammen mit einer Erweiterung des Befehlssatzes um eine Kennung, die die erwartete temporale Beziehung der Operanden einer Operation beschreibt, wodurch die Hardware verlorengegangene Datenaktualisierungen und Synchronisationsfehler erkennen kann,

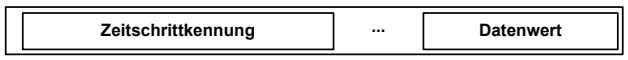



Abb. 6: Zeitschrittkennung im Datenspeicherelement

- eine Fristkennung [WH17b] (Abb. 7), die es gestattet, den Gültigkeitszeitraum der Daten einzugrenzen, wodurch die Hardware die Verwendung von Daten außerhalb dieses Zeitraums als Fehler erkennen kann,

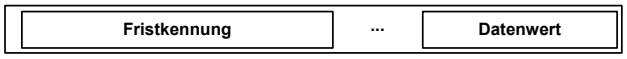



Abb. 7: Fristkennung im Datenspeicherelement

- eine Zykluszeitkennung [WH17g] (Abb. 8), die beschreibt, innerhalb welcher zeitlicher Grenzen eine aktualisierte Version eines Datenwerts erwartet wird, wodurch die Hardware ausbleibende und verfrühte Aktualisierungen des betreffenden Datenwerts als Fehler erkennen kann,

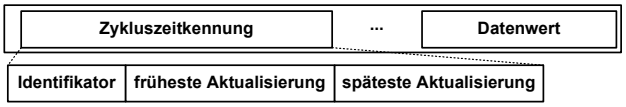



Abb. 8: Zykluszeitkennung im Datenspeicherelement

- eine Signaturkennung [WH17d] (Abb. 9) für besonders anspruchsvolle Anwendungen wie Chipkarten, bei denen der hohe Laufzeitaufwand für Prüfung und Erstellung einer kryptographischen Signatur jedes einzelnen Datenspeicherelements zur Sicherstellung der Schutzziele Integrität und Authentizität zu rechtfertigen ist,

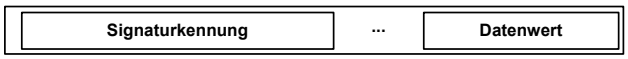



Abb. 9: Signaturkennung im Datenspeicherelement

- Datenportale in Form von Dateneingangs- und -ausgangsportalen, die es ermöglichen, Daten mit Einbeziehung der Speicheradresse in die Integritätsprüfung bzw. Signatur zwischen Systemkomponenten zu übertragen, wobei die Dateneingangsportale bei Nutzung einer Signaturkennung zusätzlich eine Prüfung der Signatur und die Umsignierung auf den eigenen geheimen Schlüssel durchführen, und
- die Vorstellung einer Realisierungsmöglichkeit der Merkmale einer Datenspezifikationsarchitektur in Datenflussarchitekturen durch Erweiterung der Verarbeitungseinheiten, um auch in diesen eine Vielzahl der identifizierten Fehler- und Angriffsarten erkennen zu können.

Unter Nutzung aller in der Arbeit vorgestellten Kennungen ergibt sich für alle innerhalb einer Datenspezifikationsarchitektur genutzten Datenspeicherelemente der in Abbildung 10 gezeigte atomare Aufbau.

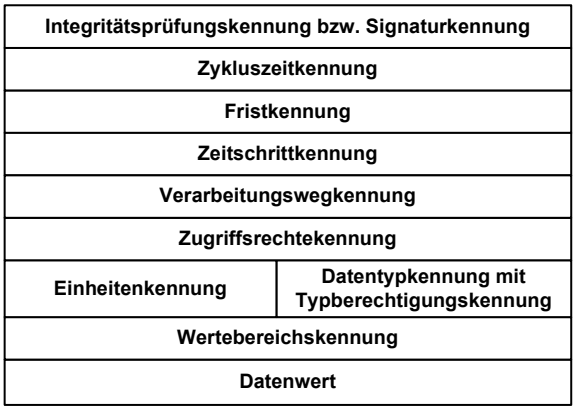



Abb. 10: Aufbau der Datenspeicherelemente einer DSA

Durch die hardwarebasierte Prüfung der in den Kennungen spezifizierten Dateneigenschaften ist es der Datenspezifikationsarchitektur möglich, alle 20 Fehler- und Angriffsarten zur Laufzeit zu erkennen. Dies wird in Tabelle 1 ersichtlich, in der die neue Architektur den konventionellen Architekturen x86 und ARM gegenübergestellt wird. Dabei wird ihre Überlegenheit in Bezug auf die Fehler- bzw. Angriffserkennung deutlich. Die meisten der Prüfungen können durch die Hardware parallel zur Ausführung der eigentlichen Operationen stattfinden, wodurch sich der Laufzeitaufwand nicht erhöht.

Durch den Einsatz der vorgestellten Fehlererkennungsmerkmale einer Datenspezifikationsarchitektur hätten in den eingangs vorgestellten Praxisbeispielen Ariane 5, Mars Climate Orbiter und Therac 25 die ursächlichen Fehler entweder vermieden oder deren Auswirkungen durch deren Erkennung und angemessene Behandlung deutlich reduziert werden können.

Tab. 1: Fehlererkennung durch konventionelle Architekturen und die Datenspezifikationsarchitektur

| **Fehlerart** | **x86, ARM** | **DSA** | **Merkmal der DSA** |
|---|---|---|---|
| Inkompatible Datentypen | – | + | Datentypkennung |
| Inkompatible Einheiten | – | + | Einheitenkennung |
| Wertebereichsunter- bzw. -überschreitung | ◯ x86 | + | Wertebereichskennung |
| Genauigkeitsproblem | – | + | Messwertdatentypen mit Werteintervall |
| Falsche Operandenauswahl | – | + | Speicheradresse in Integritätsprüfungs- oder Signaturkennung |
| Falsche Operatorauswahl | – | + | Diversitäre ALE |
| Fehlerhaftes Operationsergebnis | – | + | Diversitäre ALE |
| Fristüberschreitung | – | + | Fristkennung |
| Zyklusunterschreitung | – | + | Zykluszeitkennung |
| Zyklusüberschreitung | – | + | Zykluszeitkennung |
| Verlorengegangene Datenaktualisierung | – | + | Zeitschrittkennung |
| Synchronisationsfehler oder unvollständige Datenübertragung | – | + | Zeitschrittkennung |
| Pufferunter- oder -überläufe | (+) x86 | + | Sichere Felder, Datentypkennung |
| Fehlerhafter Datenfluss (falsche Adressaten, ...) | – | + | Verarbeitungswegkennung |
| Duplizierte Daten | – | + | Zeitschrittkennung |
| Durch Fehler oder Störungen verfälschte Daten | ◯ | + | Integritäts- oder Signaturkennung |
| Fehlerhafter Datenzugriff (fehlende Zugriffsrechte) | ◯ | + | Zugriffsrechtekennung |
| Nutzung nicht initialisierter Daten | – | + | Zugriffsrechtekennung |
| **Angriffsart** | | | |
| Gezielt verfälschte Daten | – | + | Signaturkennung |
| Wiedereinspielungsattacke | – | + | Signaturkennung mit Zeitschritt-, Frist- und Zykluszeitkennung |

Fehlererkennung: – nicht möglich, ◯ begrenzt möglich, (+) mit Einschränkungen möglich, + möglich;
DSA: Datenspezifikationsarchitektur; ALE: Arithmetisch-Logische Einheit

## Literaturverzeichnis

[AE] AEG Datenverarbeitung: TR 4 Bedienungshandbuch.

[AM12] AMD: AMD64 Architecture Programmer's Manual Vol. 2: System Programming. `http://developer.amd.com/wordpress/media/2012/10/24593_APM_v2.pdf`, 2012.

[AR11] ARM Limited: Migrating from IA-32 to ARM. Application Note 274, ARM DAI 0274, 2011.

[Br60] Brown, D. T.: Error Detecting and Correcting Binary Codes for Arithmetic Operations. IRE Transactions on Electronic Computers, Vol. EC-9, Issue 3, 1960.

[Fe72] Feustel, E.: The Rice research computer: a tagged architecture. AFIPS '72 (Spring) Proceedings of the May 16-18, 1972, spring joint computer conference, S. 369–377, 1972.

[Fe73] Feustel, E.: On the Advantages of Tagged Architectures. IEEE Transactions on Computers, Vol. C-22, No. 7:644–656, 1973.

[Fo89] Forin, P.: Vital Coded Microprocessor Principles and Application for Various Transit Systems. IFAC Control, Computers, Communications, S. 79–84, 1989.

[Gi93] Giloi, W.: Rechnerarchitektur. Springer-Verlag, 2. Auflage, 1993.

[Go14] Gollub, L.: Verfahren zur Kontrollflussüberwachung in sicherheitsgerichteten Rechensystemen. VDI Verlag GmbH, 2014.

[IE16] IEC 61784-3: Industrial communication networks - Profiles - Part 3: Functional safety fieldbuses - General rules and profile definitions. 2016.

[In91] Intel: 80386 System Software Writer's Guide. 1991.

[Ł95] Łent, B.: A Contribution To The Design Of A Disjunctive Computer Architecture For Real Time Control Systems. Dissertation, FernUniversität in Hagen, 1995.

[Li96] Lions, J.-L. et al.: Ariane 501 Inquiry Board report. `http://esamultimedia.esa.int/docs/esa-x-1819eng.pdf`, 1996.

[LT93] Leveson, N. G.; Turner, C. S.: An Investigation of the Therac-25 Accidents. Computer, Vol. 26, Issue 7:18–41, 1993.

[MS07] Meixner, A.; Sorin, D. J.: Error Detection Using Dynamic Dataflow Verification. 16th International Conference on Parallel Architecture and Compilation Techniques (PACT 2007), S. 104–118, 2007.

[Or14] Oracle: Introduction to SPARC M7 and Application Data Integrity (ADI). `https://swisdev.oracle.com/_files/What-Is-ADI.html`, 2014.

[RF81] RFC 793: Transmission Control Protocol. DARPA Internet Program, Protocol Specification, 1981.

[Sc11] Schiffel, U.: Hardware Error Detection Using AN-Codes. Dissertation, Technische Universität Dresden, 2011.

[St99] Stephenson, A. G. et al.: Mars Climate Orbiter Mishap Investigation Board Phase I Report. `ftp://ftp.hq.nasa.gov/pub/pao/reports/1999/MCO_report.pdf`, 1999.

[Te87] Teller, J.: Problematik der Datenflussfehler. Angewandte Informatik, Ausgabe 29, Nr. 6:240–247, 1987.

[Te15] Texas Instrument: Safety Manual for RM48x Hercules ARM-Based Safety Critical Microcontrollers. User's Guide, Dokumentennummer SPNU577D, 2015.

[WH17a] Widmann, S.; Halang, W. A.: Vorrichtung und Verfahren zur gerätetechnischen Einschränkung der zulässigen Operationen auf Daten in Datenverarbeitungseinheiten. Patentanmeldung 10 2017 005 945.4 beim Deutschen Patent- und Markenamt, 2017.

[WH17b] Widmann, S.; Halang, W. A.: Vorrichtung und Verfahren zur gerätetechnischen Erkennung der Datennutzung außerhalb ihres Gültigkeitszeitraums in Datenverarbeitungseinheiten. Patentanmeldung 10 2017 005 974.8 beim Deutschen Patent- und Markenamt, 2017.

[WH17c] Widmann, S.; Halang, W. A.: Vorrichtung und Verfahren zur gerätetechnischen Erkennung inkompatibler Operandeneinheiten in Datenverarbeitungseinheiten. Patentanmeldung 10 2017 005 975.6 beim Deutschen Patent- und Markenamt, 2017.

[WH17d] Widmann, S.; Halang, W. A.: Vorrichtung und Verfahren zur gerätetechnischen Erkennung von absichtlichen oder durch Störungen und / oder Fehler verursachten Datenverfälschungen in Datenverarbeitungseinheiten. Patentanmeldung 10 2017 006 354.0 beim Deutschen Patent- und Markenamt, 2017.

[WH17e] Widmann, S.; Halang, W. A.: Vorrichtung und Verfahren zur gerätetechnischen Erkennung von Datenflussfehlern in Datenverarbeitungseinheiten und -systemen. Patentanmeldung 10 2017 005 972.1 beim Deutschen Patent- und Markenamt, 2017.

[WH17f] Widmann, S.; Halang, W. A.: Vorrichtung und Verfahren zur gerätetechnischen Erkennung von Synchronisations- und Datenaktualisierungsfehlern in Datenverarbeitungseinheiten. Patentanmeldung 10 2017 005 970.5 beim Deutschen Patent- und Markenamt, 2017.

[WH17g] Widmann, S.; Halang, W. A.: Vorrichtung und Verfahren zur gerätetechnischen Erkennung von Verletzungen von zyklischen Echtzeitbedingungen in Datenverarbeitungseinheiten und -systemen. Patentanmeldung 10 2017 005 944.6 beim Deutschen Patent- und Markenamt, 2017.

[WH17h] Widmann, S.; Halang, W. A.: Vorrichtung und Verfahren zur gerätetechnischen Erkennung von Wertebereichsverletzungen von Datenwerten in Datenverarbeitungseinheiten. Patentanmeldung 10 2017 005 971.3 beim Deutschen Patent- und Markenamt, 2017.

[Wi17] Widmann, S.: Eine Datenspezifikationsarchitektur. VDI Verlag GmbH, 2017.

**Stefan Widmann** wurde in Schwäbisch Gmünd geboren und studierte nach dem Abitur Industrieelektronik an der Fachhochschule Ulm. Nach Erlangung seines Diploms im Jahr 2004 arbeitete er zunächst bei einem Mittelständler als Entwicklungsingenieur im Bereich Leistungselektronik. 2006 wechselte er zur Siemens AG in Amberg, wo er seither als Projektleiter, Firmwarearchitekt und -entwickler in der Leistungsschalterentwicklung tätig ist. Seit 2008 betreibt er zudem als Freiberufler ein eigenes Ingenieurbüro. Nebenberuflich absolvierte er von 2010 bis 2014 das Masterstudium der Elektro- und Informationstechnik an der FernUniversität in Hagen und wurde dort anschließend unter der Betreuung von Prof. Dr. Dr. W. A. Halang im Jahr 2017 zum Doktor-Ingenieur promoviert. Für seinen Masterabschluss und seine Dissertation wurde er jeweils mit einem Preis ausgezeichnet. Sowohl seine Masterarbeit, als auch seine Dissertation sind als Bücher im VDI Verlag erschienen.

# Anisotrope Röntgendunkelfeldtomographie[1]

Matthias Wieczorek[2]

**Abstract:** Moderne Röntgen-Bildgebung ermöglicht Aufnahmen von Phasenkontrast- (Brechung) und Dunkelfeld-Informationen (Streuung). Die Rekonstruktion des Dunkelfeldsignals stellt ein besonders anspruchsvolles Problem dar, da die Streuung innerhalb eines Objektes von dessen Orientierung abhängt. In dieser Arbeit wird sowohl ein abstraktes Software-Framework für tomographische Rekonstruktion als auch eine neuartige Methode zur Anisotropen Röntgen-Dunkelfeld Tomographie vorgestellt. Ein erstes biomedizinisches Experiment an einer Probe eines menschlichen Cerebellums weist daraufhin, dass diese Methode eine komplementäre bildgebende Methode zur Abbildung von Nervenfasern liefern könnte.

## 1 Einleitung

Eine der bedeutendsten Entdeckungen für die moderne Medizin war die der Röntgenstrahlen durch Wilhelm Conrad Röntgen [Rö96]. Für diese Arbeit erhielt er 1901 den allererste Nobel Preis. Die essentielle Eigenschaft dieser hochenergetischen Form der Strahlung ist, dass sie Objekte durchdringen kann, welche für andere elektromagnetische Strahlung wie zum Bespiel dem sichtbaren Lichtspektrum undurchsichtig sind. Ein Röntgenstrahl der ein Objekt durchläuft wird durch spezifische physikalische Eigenschaften dieses Objektes verändert. Das besondere ist nun, dass der Röntgenstrahl Information entlang des kompletten Pfades durch das Objekt sammelt. Im Anschluss kann das akkumulierte Ergebnis dieser Veränderungen auf dem Röntgenbild abgelesen werden. Im Fall der klassischen Röntgenbildgebung handelt es sich hierbei um die Absorption, d.h. ein Teil der Photonen des Röntgenstrahls werden von dem gemessenen Objekt absorbiert. Die Absorption an einer bestimmten Stelle innerhalb eines Objektes hängt mir der Dichte der zu durchdringenden Materie an dieser Stelle zusammen, sodass beim Menschen insbesondere harte Strukturen wie zum Beispiel Knochen gut sichtbar sind. Weiches Gewebe hingegen zeigt nur einen schwachen Bildkontrast.

Bis zum Beginn dieses Jahrhunderts war die Absorption die einzige Information, welche man mit klinischen Röntgengeräten messen konnte. Erst innerhalb der letzten 15 Jahre wurden Methoden entwickelt, um weitere Einflüsse, denen der Röntgenstrahl naturgemäß unterliegt, sichtbar beziehungsweise messbar zu machen. Im Folgenden soll ein kurzer Überblick über absorptionsbasierte Verfahren geben werden, um im Anschluss meine Arbeit [Wi17a] in das Feld der Röntgenbildgebung einzubetten.

[1] Englischer Titel der Dissertation: “Anisotropic X-ray Dark-field Tomography”

[2] Technische Universität München, wieczore@in.tum.de

## 2 Röntgenbildgebung und Computertomographie

Die absorptionsbasierten Röntgenbildgebung ermöglichte es erstmals einen Blick in das Innere von Menschen zu werfen, ohne dass man dazu den Patienten aufschneiden musste. Dasselbe gilt für die Untersuchung von nicht organischen Objekten im Rahmen der zerstörungsfreien Prüfung. Ein Problem bei der klassischen Röntgenbildgebung ist jedoch, dass das resultierende Bild, alle Effekte akkumuliert abbildet, weswegen einzelne Strukturen entlang des Strahls nicht isoliert betrachtet werden können. Wichtige Tiefeninformationen gehen dadurch verloren und eine Interpretation ist nur durch ein spezielles Training und Wissen über die Anatomie möglich. Es gab mehrere Ansätze mittels mehreren Röntgenaufnahmen Rückschlüsse auf die jeweilige Absorption an jeder Stelle des gemessenen Objektes zu schliessen um diese verlorengegangene Tiefeninformation wieder herzustellen.

Dies gelang in den 60er Jahren Sir Godfried Hounsfield [Ho73] und Allan M. Cormack [Co63, Co64] mit der Erfindung der Computertomographie (CT). Bei diesem Verfahren werden von einem Objekt oder einem Patienten eine Vielzahl von einzelnen Röntgenaufnahmen auf einer kreis- oder einer helixförmigen Bahn gemacht. Aus den Bildern der unterschiedlichen Perspektiven wird nun der physikalische Effekt, d.h. die Absorption, an jeder Position des Körpers berechnet. Dies bezeichnet man als tomographischen Rekonstruktion. Der essentielle Bestandteil der für eine solche Tomographie benötigt wird ist ein sogenanntes Vorwärtsmodel. Dies ist ein mathematisches Model, welches Röntgenbilder für eine gegebenen Absorptionsverteilung in einem Körper simuliert. Zur Berechnung der tomographische Rekonstruktion, muss man dieses Vorwärtsmodel nun invertieren, um aus den zweidimensionalen Röntgenbildern das dreidimensionale Objekt wieder herzustellen. Das Ergebnis einer Computertomographie ist eine dreidimensionale Repräsentation der Absorption innerhalb des gemessenen Objektes. Bei klassischem CT erhält man erneut besonders gute Kontraste für harte Materialien, während Unterschiede in weichem Gewebe deutlich schlechter sichtbar sind.

Insbesondere um einen Einblick in weiches Gewebe zu erhalten wurden daher dauerhaft Bemühungen betrieben bildgebende Verfahren zu entwickeln, welche weitere Strahlungsarten oder -modalitäten nutzen um zusätzliche oder bessere Sichtbarkeit zu erzielen. Einige der wohl bekanntesten Verfahren sind die Ultraschallbildgebung (US) und die sogenannte Magnetresonanztomographie (MRT).

### 2.1 Motivation

Genau diese Motivation ist auch der Ausgangspunkt für die von mir verfasste Dissertation [Wi17a] gewesen. Während wir im Gebiet der Röntgenstrahlung verbleiben, untersuchen wir einen anderen physikalischen Einfluss auf diese.

Hierbei nutzt man, dass die Absorption nicht die einzige Auswirkung eines Objektes auf einen Röntgenstrahl ist, sondern, wie man es auch von sichtbarem Licht kennt, der Röntgenstrahl ebenfalls gebrochen und/oder gestreut wird. Das Problem ist jedoch, dass

die hohe Energie, die auf der einen Seite ermöglicht, dass Röntgenstrahlung z.B. einen menschlichen Körper durchdringen kann, auf der anderen Seite dazu führt das die Auswirkungen von Streuung und Brechung sehr klein sind. Lange Zeit stellte dies ein erhebliches Problem dar und lediglich an speziellen Einrichtungen, sogenannten Synchrotronen konnten derartige Effekte gemessen und untersucht werden. Für Röntgenmessgeräte, wie sie z.B. im klinischen Alltag verwendet werden, waren diese Effekte der Brechung und Streuung jedoch zu klein, so dass sie nicht von einem Röntgendetektor aufgelöst werden konnten. Dies bedeutet aber auch dass die zugrundeliegenden Ursachen ebenfalls deutlich kleiner (im Mikrometerbereich) sind, als die Auflösung eines Detektors.

## 2.2 Phasenkontrast und Dunkelfeld

Der Durchbruch gelang erst vor rund 15 Jahren mit der Entwicklung von Talbot-Lau Interferometrie basierten Systemen [Mo03, We05, Pf06, Pf08].

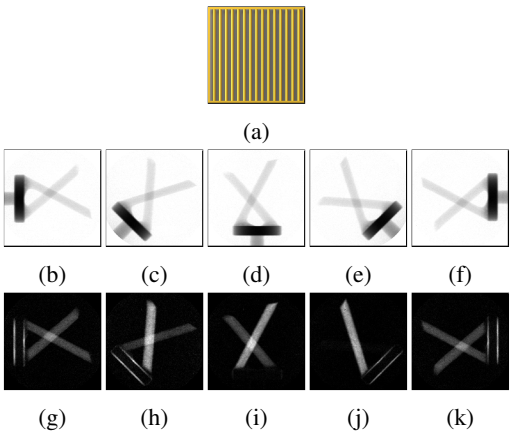



Abb. 1

Hierbei wird ein klassischer Röntgenaufbau, bestehend aus Röntgenquelle und Detektor um drei Interferometriegitter erweitert. Wird nun das mittlere dieser Gitter während der Röntgenmessung bewegt, so wird die Brechung und die Streuung in eine Veränderung der Bildwerte übersetzt. Man unterscheidet hierbei den sogenannten Phasenkontrast (c.f. [Pf06]) (Brechung) und dem Dunkelfeld (c.f. [Pf08]) (Kleinwinkelröntgen-streuung) . Damit ist dieser Aufbau in der Lage die Information über Brechung und Streuung so zu übersetzt, dass ein normaler Röntgendetektor diese abbilden kann. Das Besondere hierbei ist, dass dies nun für Bereiche in denen die absorptionbasierte Röntgenbildgebung nur schlechten bis keinen Kontrast liefert, deutlich bessere und zusätzliche Information liefert. Zudem geben die resultierenden Bilder, Informationen über Mikrostrukturen innerhalb des Objektes wieder, welche deutlich kleiner als die Auflösung des Detektors sind. Von besonderem Interesse für die von mir verfasste Dissertation ist das Dunkelfeldsignal. Analog zu klassischem CT stellt sich die Frage, ob es möglich ist, ebenfalls die zugrundeliegenden physikalischen Effekte an jeder Stelle des gemessenen Objektes zu berechnen. Bei klassischer Computertomographie lässt sich die Absorption an jeder Position innerhalb des gemessenen Objektes durch einen einzelnen Wert darstellen. Diese Werte bezeichnet man als lineare Attenuationskoeffizienten. Dies gilt auch im Falle des Phasenkontrastes. Es handelt sich also um isotrope Eigenschaften. Für das Dunkelfeld hingegen, gilt dies nicht. Für faserartige Strukturen ist die Röntgenstreuung rechtwinklig zu den faserartigen Strukturen am stärksten (c.f. [Je10b, Je10a]). Damit enthält dieses Signal nicht nur Informationen über die Stärke des

physikalischen Effektes der Streuung, der Streustärke, sondern zusätzlich über die Richtung der Streuung. Dadurch stellt die tomographische Rekonstruktion des Dunkelfeldsignals ein besonders anspruchsvolles Problem dar, da die Streustärke, im Gegensatz zu der Absorption und dem Phasenkontrast, von der relativen Orientierung einer Mikrostruktur zu den Gittern des Aufbaus abhängt. Da die Gitter in dem Talbot-Lau Interferometrie basierten System die Streustärke, welche orthogonal zu den Gitterstäben stattfindet, messen (c.f. [Ma13]). Man spricht hierbei von einem anisotropen Signal.

Der Effekt wird auf Abb. 1 verdeutlicht. Zu sehen sind zwei Holzstöckchen. Holz hat von Natur aus Mikrofaserstrukturen in Wachstumsrichtung. Wird nun diese Probe in der Ebene parallel zu den Gittern gedreht, so zeigt sich, dass sich die Intensität in der Absorption (Bilder (b) bis (f)) nicht ändert, wohingegen das Dunkelfeldsignal am stärksten ist, wenn einer der Stöckchen mit den Gitterstäben ausgerichtet ist. Erneut stellt das resultierende Bild lediglich die aufsummierten Einflüsse entlang des Röntgenstrahls dar und es stellt sich die Frage nach der tomographischen Rekonstruktion des Streuprofils an jeder Position des gemessenen Objektes.

Die Anisotropie des Dunkelfeldsignals stellt jedoch nicht nur bei der Rekonstruktion eine Herausforderung dar, sondern es stellt sich bereits bei der Aufnahme der Dunkelfeldbilder eine Besonderheit heraus. Nämlich, dass eine einfache Rotation des Aufbaus um das Objekt nicht ausreicht um diese Information ausreichend zu erfassen (c.f. [Ma14]).

## 3 Röntgentensortomographie

Eine erste Methode um sowohl die nötigen Dunkelfeldbilder aufzunehmen, als auch eine tomographische Rekonstruktion zu berechnen wurde von Malecki et al. entwickelt [Ma14]. Diese Methode wird als X-Ray Tensor Tomography (XTT) (deutsch: Röntgentensortomographie) bezeichnet. Für diese Methode wird ein sogenannter Rang-2-Tensor verwendet um die Streuung in jeder Position des Messobjektes zu beschreiben. Diesen Tensor kann man sich als Ellipsoid. Dieses Ellipsoid orientiert sich rechtwinklig zu der Mikrofaserstruktur. Dieser Tensor vereint Informationen über die Streustärke sowie die Richtungsverteilung, welche einen Einblick in die Orientierung der Mikrostrukturen innerhalb des Objektes liefern. Eine wesentliche Einschränkung des XTT Ansatzes liegt jedoch darin, dass ein Tensor auf eine einzige Mikrostrukturrichtung beschränkt ist (c.f. [Wi16]). Da wir jedoch Mikrostrukturen messen, die deutlich kleiner als die Auflösung unseres Systems liegen vermuteten wir, dass mehrere Fasern unterschiedlicher Richtungen dasselbe Volumenelement kreuzen können. Es stellt sich also die Frage nach einer vollständiger Beschreibung der Streuung in einem dieser Volumenelemente, sodass auch mehrere Richtungen beschrieben werden können.

## 4 Anisotrope Röntgendunkelfeldtomographie

Das Grundprinzip bei XTT [Ma14] besteht darin, dass zu Beginn der tomographischen Rekonstruktion eine beliebige Anzahl an Streurichtungen heraussucht wird und man annimmt, dass Streuung nur in diese Richtungen vorherrscht. Mit Hilfe dieser Annahme

lässt sich eine anisotrope Dunkelfeldaufnahme dadurch simulieren, dass an jeder Stelle innerhalb des Objektes die Streuung bezüglich der Streustärke in dieser Richtungen, der relativen Orientierung der Richtungen und der Gitterstäbe und zu guter Letzt der Röntgenstrahlrichtung, aufsummiert wird [Ma13]. Die resultierenden Werte an jeder Stelle im Objekt werden nun, analog zur klassischen Computertomographie entlang eines Röntgenstrahls aufsummiert.

Bei der Rekonstruktion von XTT wird dieses Model nun invertiert [Ma13, Vo15]. Hierbei werden die individuellen Streustärken bezüglich der zuvor gewählten Richtungen rekonstruiert. Diese Information wird dann vereint durch einen Tensor dargestellt. Hierbei hängt die Qualität der Rekonstruktion sehr stark von den gewählten Streurichtungen ab. Umso mehr und feiner man diese Streurichtungen wählt, umso genauer und feiner werden auch die rekonstruierten Richtungsinformationen über die Mikrofasern (see [Wi17a]). Auf der anderen Seite wird aber auch die Berechnung entsprechend aufwändiger, was sowohl erhöhten Rechen- als auch Zeitaufwand mit sich bringt.

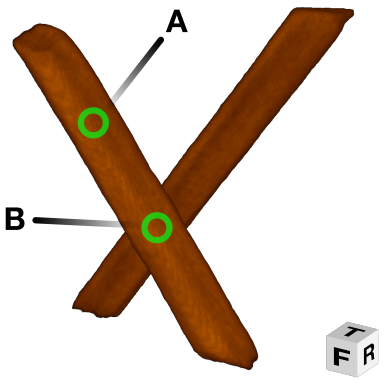
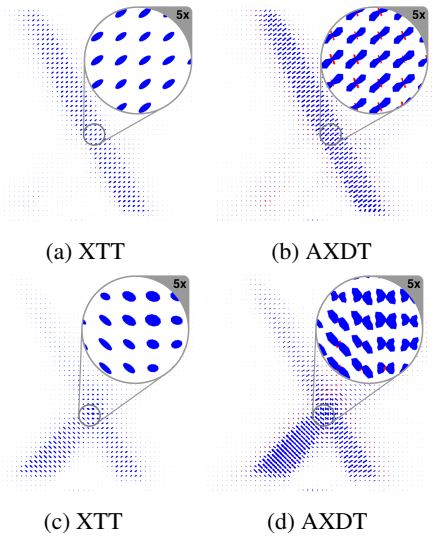



Abb. 2: Aus M. Wieczorek, F. Schaff, F. Pfeiffer, und T. Lasser. "Anisotropic X-Ray Dark-Field Tomography: A Continuous Model and its Discretization". In: *Physical Review Letters* 117.15 (Okt. 2016), p. 158101, mit der Erlaubnis durch die APS (©2016 American Physical Society).

Anstelle dieser diskreten Streurichtungen stellen wir nun die Streuung an jeder Position innerhalb der Objektes durch eine sphärische Funktion dar. Das bedeutet, dass jeder beliebigen Streurichtung eine Stärke zugewiesen wird, anstatt nur die diskreten Richtungen zu betrachten. Diese spezielle Familie an Funktionen durch sogenannte Kugelflächenfunktionen dargestellt werden können. Daher waren wir in der Lage die Streufunktion an jeder Position durch Kugelflächenfunktionen dar. Dies ermögliche es uns ein Vorwärtsmodel für Dunkelfeldtomographie direkt auf Basis der entsprechenden Koeffizienten der Kugelflächenfunktionen zu postulieren [Wi16]. Dieses neue Vorwärtsmodel und die daraus resultierende tomographische Rekonstruktion nennen wir anisotrope Röntgen-dunkelfeldtomographie (engl. AXDT). Die Rekonstruktion berechnet nun die jeweiligen Beiträge der entsprechenden Kugelflächenfunktionen an jeder Stelle des gemessenen Objektes. Darüber hinaus konnten wir zeigen, dass es mit unserem aktuellen Verständnis des Dunkelfeldbildgebung ausreicht die ersten 15 Kugelflächenfunktionen zu betrachten um eine exakte Rekonstruktion zu erhalten [Wi16]. Da die Komplexität der Berechnung einer Rekonstruktion gleich zu der bei XTT ist, bedeutet dies, dass wir eine exakte Rekonstruktion berechnen können, mit dem gleichen Aufwand, den wir bei XTT benötigen um lediglich 15 Streurichtungen zu berechnen.

Weiter stellte sich nun die Frage ob AXDT es uns ermöglicht mehrere Mikrostrukturrichtungen innerhalb eines Volumenelements zu erkennen. Hierzu haben wir erneut die Probe mit den zwei gekreuzten Stöckchen verwendet. Unsere Hypothese war, dass wir hierbei Volumenelemente identifizieren können, durch die Fasern mit unterschiedlichen Richtungen verlaufen. Zu diesem Zweck haben wir zwei Bereiche innerhalb der Rekonstruktion von dieser Probe untersucht (Abb. 2).




(a) XTT (b) AXDT

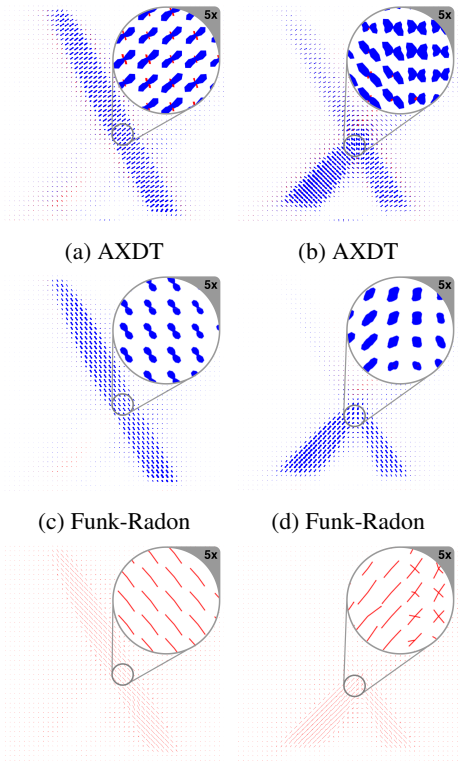
(c) XTT (d) AXDT

Abb. 3: M. Wieczorek, F. Schaff, F. Pfeiffer, und T. Lasser. "Anisotropic X-Ray Dark-Field Tomography: A Continuous Model and its Discretization". In: *Physical Review Letters* 117.15 (Okt. 2016), p. 158101, mit der Erlaubnis durch die APS (©2016 American Physical Society).

Zum einen, einen Schnitt durch die Rekonstruktion, in welchem sich lediglich einer der beiden Stöcke befindet und zum anderen einen Schnitt durch die Ebene, in der sich die beiden Stöcke berühren. Unsere Vermutung war, dass wir innerhalb der Ebene in der sich die Stöcke berühren, Stellen finden, in denen Strukturinformationen von beiden Stöcken zu finden sind [Wi16].

Die Visualisierung der Ergebnisse sind in der Abb. 3. Wir finden in der ersten Ebene (a), (b) Streuung welche orthogonal zu der Wachstumsrichtung des Stöckchens ist, was einer Mikrostrukturrichtung entspricht. Die Ergebnisse von XTT und AXDT weisen auf die selbe Richtung hin. Dies weist darauf hin, dass wir mit der hier vorgestellten AXDT Methode nach wie vor gleiche Strukturinformation erhalten, falls nur eine Richtung vorherrscht. Interessanter ist jedoch die Ebene, in der sich die beiden Stöckchen berühren (c), (d). Hier zeigt sich in den XTT Ergebnissen kein Hinweis auf mehrere Richtungen, da der Tensor dies nicht ausdrücken kann. Vielmehr führt die Tensorbeschreibung dazu, dass die Information gemittelt wird, sodass keine der beiden Richtungen dargestellt wird. Im Gegensatz dazu, zeigen sich bei der AXDT Methode Streuprofile, welche auf zwei Richtungen hinweisen, jeweils eine orthogonal zu einem der beiden Stöckchen. Diese Ergebnisse deuten damit darauf hin, dass wir bei nahezu gleichem Rechenaufwand in der Lage sind mit AXDT mehrere Streurichtungen zu rekonstruieren. Diese Ergebnisse wurden im Rahmen der folgenden Publikation präsentiert: [Wi16].

Wie in den Ergebnissen oben zu erkennen ist, ist die Streuung orthogonal zu Mikrostrukturen am stärksten. Im Falle von XTT sind diese durch die Richtung der kleinsten Halbachse des Tensors gegeben (c.f. [Vo15, Ma13]). Diese Analogie funktioniert nicht mehr, wenn man mehrere Richtungen innerhalb eines Volumenelements hat.




(a) AXDT (b) AXDT

(c) Funk-Radon (d) Funk-Radon

(e) Richtung (f) Richtung

Abb. 4: (a),(b) sind aus M. Wieczorek, F. Schaff, F. Pfeiffer, und T. Lasser. "Anisotropic X-Ray Dark-Field Tomography: A Continuous Model and its Discretization". In: *Physical Review Letters* 117.15 (Okt. 2016), p. 158101, mit der Erlaubnis durch die APS (©2016 American Physical Society).

Daher stellte sich die Frage, ob man aus den so rekonstruierten Streuprofilen diese Mikrostrukturrichtungen extrahieren kann. Es zeigte sich das die sogenannte Funk-Radon Transformation ([Fu13]) genau dies ermöglicht [WPL17]. Sie bildet Maxima entlang von Kreisen auf der Sphäre in die Richtung orthogonal zu diesem Kreis ab. Diese Transformation kann darüber hinaus besonders effiziente berechnet werden im Falle der von uns verwendeten Kugelflächenfunktionen [Fu13]. Um die Mikrostrukturrichtungen aus unserer AXDT Rekonstruktion zu erhalten berechneten wir also die Funk-Radon Transformation des rekonstruierten Streuprofils an jeder Stelle in dem Rekonstruktionsvolumen (Abb. 4) [WPL17, Wi17a].

Erneut nutzen wir die zwei Schnittebenen von oben. Die Schnittebene mit lediglich einer Mikrostrukturrichtung ist in der linken Spalte dargestellt. Die Ergebnisse von der zweiten Schnittebene, in der sich die beiden Stöckchen berühren, sind in der rechten Spalte dargestellt. Wir sehen, dass die Funk-Radon Transformation erfolgreich die Streuprofile in die jeweils orthogonale Richtung überträgt. Eine Maximumsdetektion auf diesen transformierten Streuprofilen ermöglicht nun die Extraktion der Mikrostrukturrichtungen. Diese Methode wurde im Rahmen von [WPL17] vorgestellt.

### 4.1 Biomedizinischer Anwendungsfall

In der Vergangenheit und insbesondere mit der XTT Methode wurden hauptsächlich Knochen [Sc14] und Zähne [Ju16] untersucht. Ein weiterer Bereich des menschlichen Körpers, welcher anisotrope Strukturen aufweist ist das Zentralnervensystem.

Deutlich vereinfacht besteht das menschliche Gehirn aus Neuronen, welche durch Nervenfasern verbunden sind. Da diese Fasern in der Mikrometer Größen-ordnung zu finden sind, sind diese jedoch mittels der herkömmlichen Computertomographie nicht sichtbar.

Eine Möglichkeit diese Verbindungen innerhalb des Gehirns dennoch sichtbar zu machen ist die sogenannte Diffusionsgewichtete Magnetresonanztomografie [BML94].

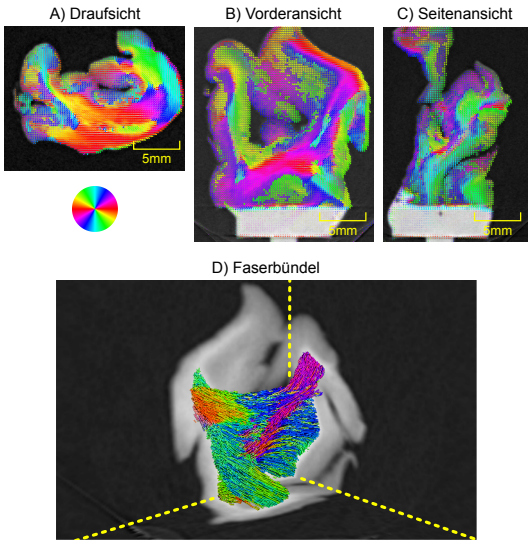



Abb. 5: Dieses Bild wurde von Matthias Wieczorek erstellt und ist lizensiert durch die Creative Commons Attribution-Noncommercial-ShareAlike 4.0 Lizenz (`https://creativecommons.org/licenses/by-sa/4.0/`).
Die ursprüngliche Version aus [Wi17a] wurde aus dem Englischen in das Deutsche übersetzt und blieb ansonsten unverändert.

In einer ersten Machbarkeitsstudie führten wir daher ein biomedizinisches Experiment an einer Probe eines menschlichen Cerebellums (Kleinhirn) durch [Wi17b]. Die resultierenden Dunkelfeldaufnahmen wurden mit dem vorgestellten AXDT Verfahren [Wi16] rekonstruiert. Anschliessend wurden die Faserstrukturrichtungen extrahiert [WPL17].

Die Visualisierung der Ergebnisse ist in Abb. 5 dargestellt. Hierbei haben wir die klassische absorptionsbasierte Computertomographie mit den Strukturinformationen der AXDT überlagert. Die Fasern entsprechen ihrer Orientierung eingefärbt. Die rekonstruierten Mikrofaserstrukturen schmiegen sich insbesondere in die sogenannte Weiße Substanz ein. Sowohl die Richtungen als auch die Lokalisierung sind plausibel und decken sich mit dem aktuellen Verständnis über das menschliche Gehirn. Damit deuten die Ergebnisse darauf hin, dass dieses Verfahren in der Lage ist als komplementäres bildgebendes Verfahren für das Zentralnervensystem zu dienen.

## Literaturverzeichnis

[BML94] Basser, P J; Mattiello, J; LeBihan, D: MR diffusion tensor spectroscopy and imaging. Biophys. J., 1994.

[Co63] Cormack, A M: Representation of a Function by Its Line Integrals, with Some Radiological Applications. J. Appl. Phys., 1963.

[Co64] Cormack, A M: Representation of a Function by Its Line Integrals, with Some Radiological Applications. II. J. Appl. Phys., 1964.

[Fu13] Funk, P: Über Flächen mit lauter geschlossenen geodätischen Linien. Math. Ann., 1913.

[Ho73] Hounsfield, G N: Computerized transverse axial scanning (tomography): Part 1. Description of system. The British Journal of Radiology, 1973.

[Je10a] Jensen, Torben H; Bech, Martin; Bunk, Oliver; Donath, Tilman; David, Christian; Feidenhans'l, Robert; Pfeiffer, Franz: Directional x-ray dark-field imaging. Phys. Med. Biol., 2010.

[Je10b] Jensen, Torben Haugaard; Bech, Martin; Zanette, Irene; Weitkamp, Timm; David, Christian; Deyhle, Hans; Rutishauser, Simon; Reznikova, Elena; Mohr, Jürgen; Feidenhans'l, Robert; Pfeiffer, Franz: Directional x-ray dark-field imaging of strongly ordered systems . Phys. Rev. B, 2010.

[Ju16] Jud, Christoph; Schaff, Florian; Zanette, Irene; Wolf, Johannes; Fehringer, Andreas; Pfeiffer, Franz: Dentinal tubules revealed with X-ray tensor tomography. Dental Materials, 2016.

[Ma13] Malecki, Andreas; Potdevin, Guillaume; Biernath, Thomas; Eggl, Elena; Garcia, Eduardo Grande; Baum, Thomas; Noël, Peter B; Bauer, Jan S; Pfeiffer, Franz: Coherent Superposition in Grating-Based Directional Dark-Field Imaging. PLOS ONE, 2013.

[Ma14] Malecki, A; Potdevin, G; Biernath, T; Eggl, E; Willer, K; Lasser, T; Maisenbacher, J; Gibmeier, J; Wanner, A; Pfeiffer, F: X-ray tensor tomography. EPL (Europhysics Letters), 2014.

[Mo03] Momose, Atsushi: Phase-sensitive imaging and phase tomography using X-ray interferometers. Opt. Express, 2003.

[Pf06] Pfeiffer, Franz; Weitkamp, Timm; Bunk, Oliver; David, Christian: Phase retrieval and differential phase-contrast imaging with low-brilliance X-ray sources. Nat. Phys., 2006.

[Pf08] Pfeiffer, F; Bech, M; Bunk, O; Kraft, P; Eikenberry, E F; Brönnimann, Ch; Grünzweig, C; David, C: Hard-X-ray dark-field imaging using a grating interferometer. Nat. Mater., 2008.

[Rö96] Röntgen, W C: Über eine neue Art von Strahlen. In: Sitzungsberichte der Physikalisch-Medizinischen Gesellschaft zu Würzburg. 1896.

[Sc14] Schaff, Florian; Malecki, Andreas; Potdevin, Guillaume; Eggl, Elena; Noël, Peter B; Baum, Thomas; Garcia, Eduardo Grande; Bauer, Jan S; Pfeiffer, Franz: Correlation of X-Ray Vector Radiography to Bone Micro-Architecture. Sci. Rep., 2014.

[Vo15] Vogel, Jakob; Schaff, Florian; Fehringer, Andreas; Jud, Christoph; Wieczorek, Matthias; Pfeiffer, Franz; Lasser, Tobias: Constrained X-ray tensor tomography reconstruction. Opt. Express, 2015.

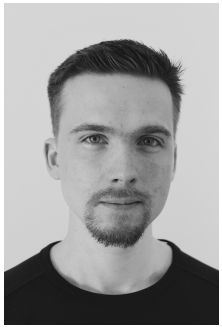
[We05] Weitkamp, Timm; Diaz, Ana; David, Christian; Pfeiffer, Franz; Stampanoni, Marco; Cloetens, Peter; Ziegler, Eric: X-ray phase imaging with a grating interferometer. Opt. Express, 2005.

[Wi16] Wieczorek, Matthias; Schaff, F; Pfeiffer, F; Lasser, T: Anisotropic X-Ray Dark-Field Tomography: A Continuous Model and its Discretization. Phys. Rev. Lett., 2016.

[Wi17a] Wieczorek, Matthias: Anisotropic X-ray Dark-field Tomography . Dissertation, 2017.

[Wi17b] Wieczorek, Matthias; Schaff, Florian; Jud, Christoph; Pfeiffer, Daniela; Pfeiffer, Franz; Lasser, Tobias: Brain connectivity exposed by Anisotropic X-ray Dark-Field Tomography: A preclinical survey. **(...)**, 2017. **(In submission.)**.

[WPL17] Wieczorek, Matthias; Pfeiffer, F; Lasser, T: Micro-structure orientation extraction for Anisotropic X-Ray Dark-Field Tomography. In: Fully3D. 2017.



**Matthias Wieczorek** wurde am 13. März 1986, in München geboren. Bereits während seinem Bachelorstudium der Informatik an der Technischen Universität München legte er seinen Schwerpunkt in dem Bereich der Medizininformatik. Nach seinem Bachelorstudiengang in Informatik begann er ein Doppelstudium, ebenfalls an der Technischen Universität München, für den Master in Informatik und einen Bachelor in Mathematik mit Nebenfach Physik. Nach dem Abschluss des Masterstudienganges begann er mit seiner Promotion im Bereich der Tomographischen Rekonstruktion am Lehrstuhl von Prof. Nassir Navab (TUM) unter der Betreuung von Dr. Tobias Lasser. Während der Zeit seiner Promotion arbeitete er eng mit dem Lehrstuhl von Prof. Pfeiffer zusammen. In einer gemeinsamen interdisziplinären Forschungsgruppe wurden Themen der moderne Röntgenbildgebung untersucht und erforscht. Am 31. Mai 2017 reichte er seine Dissertationsschrift mit dem Titel “Anisotropic X-ray Dark-field Tomography” ein, welche er am 24. November 2017 verteidigte. Direkt im Anschluss an die Einreichung trat er eine Stelle als Wissenschaftler/Softwareentwickler bei der ImFusion GmbH an, bei welcher er sich weiter mit Themen der Computertomografie beschäftigt.