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 $F := \mathrm{diag}(\mathbf{f})$ die zugehörige Diagonalmatrix. Weiterhin bezeichne $\mathbf{d} := W\mathbf{1}$ den Gradvektor (Zeilensummen) von W, somit gilt $D = \mathrm{diag}(\mathbf{d})$. 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 $Z_{\mathbf{f}} = \mathrm{diag}(\tilde{\mathbf{f}}/\mathbf{f})$ 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 Ω, 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 $\mathrm{vol}_G(S) = \sum_{i \in S} d_i$ und Graph-Schnitte $\mathrm{cut}_G(S) = \sum_{i \in S, j \notin S} w_{ij}$ eines solchen Graphen in einen Zusammenhang mit korrespondierenden Bereichsintegralen der zugrundeliegenden Dichte bringen lassen. Anders als die Knotengrade korrespondieren vol_G und 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 Ω, 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 nicht-ganzzahliger Intensivierungen von p. Weiterhin bezieht sich der $\mathbf{d}^{-1}$-adjusted Graph auf die Gleichverteilung auf Ω. Dies erlaubt es, im Graph-Clustering die rein geometrische Struktur von Ω 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 $W \in \mathbb{R}_{\geq 0}^{m \times n}$ 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_∞ 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_∞ 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_∞ 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_∞ 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_∞ 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 $EV(L) \Leftrightarrow \text{RQC}(L) \overset{relax}{\rightsquigarrow} \text{MinRatioCut}$.

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 $EV(\overline{L}) \Leftrightarrow \mathrm{RQC}(\overline{L}) \overset{relax}{\leftsquigarrow}$ 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.