Transparenz durch Privacy Dashboards: Ein Process Mining Ansatz

Christian Zimmermann, Rafael Accorsi
Business Process Security Group
Universität Freiburg
{zimmermann, accorsi}@iig.uni-freiburg.de

Abstract: Präventive Datenschutztechniken alleine reichen nicht mehr aus, um die Privatheit von Nutzern datenzentrischer Dienste zu schützen. Dieser Artikel berichtet von unserer laufenden Arbeit hinsichtlich des Designs von Privacy Dashboards und der ihnen zugrunde liegenden Infrastruktur. Der vorgestellte Ansatz zielt darauf ab, den Schutz der Privatheit durch Transparenz hinsichtlich Datenspeicherung und -verwendung durch Dienstanbieter zu gewährleisten. Unser Ansatz beruht auf der Kombination von anbieterseitigen Privacy Dashboards mit Methoden des Process Minings auf Basis vertrauenswürdiger Logdateien und Trusted Computing Platforms. Der Ansatz beruht dabei nicht auf Datenvermeidung, sondern auf Transparenz und Kontrolle preisgegebener Daten.

1 Einleitung

Moderne, *datenzentrische* [MFP12] Geschäftsmodelle des E-Business beruhen auf der Auswertung und Monetarisierung von Nutzerdaten. Datenzentrische Dienstanbieter wie etwa Google oder Facebook stellen Nutzern ihre Dienste dazu kostenfrei zur Verfügung und erwirtschaften Gewinne hauptsächlich durch personalisierte Werbeschaltungen. So haben im Jahr 2012 bspw. Google über 96% und Facebook 84% ihres Gesamtumsatzes mittels Werbeschaltungen erwirtschaftet [Goo13, Fac13]. Grundlage personalisierter Werbung sind dabei automatisiert generierte Nutzerprofile, die neben demographischen Informationen über Nutzer auch Interessenkategorien beinhalten, mittels derer sich Nutzer passgenau in Werbezielgruppen einordnen lassen. Damit kann jedem Nutzer indiviuell auf sein Profil zugeschnittene Werbung angezeigt werden.

Nutzerprofile basieren nicht nur auf von Nutzern wissentlich preisgegebenen Informationen, wie etwa Angaben von Interessen und Hobbys innerhalb eines Online Social Network Dienstes. Zusätzlich fließen auch unwissentlich preisgegebene Informationen in Nutzerprofile ein. Dazu gehören u.a. Informationen, die Nutzer durch ihr Browsingverhalten im Internet preisgeben oder Informationen, die automatisiert aus bereits preisgegebenen Daten inferiert werden können [AZM12]. So können bspw. durch die Analyse des sozialen Netzes eines Nutzers, also der Beziehungen des Nutzers zu anderen, vom Nutzer nicht direkt und wissentlich preisgegebene Informationen inferiert werden [MSLC01, YLS+11].

Angesichts der allgegenwärtigen Sammlung und Auswertung persönlicher Daten wird von

manchen bereits der Tod des auf Datenvermeidung basierenden Verständnisses von Privatheit prophezeit [Sol08]. Tatsächlich ist der Schutz persönlicher Daten gegenüber den Anbietern datenzentrischer Dienste mit herkömmlichen, präventiven, technischen Mitteln des Datenschutzes (*Privacy-Enhancing Technologies, PETs*) alleine nahezu unmöglich geworden. So sind präventive Datenschutzmechanismen konzeptionell kaum geeignet, bspw. das Inferieren von Informationen über Nutzer aus bekannten Informationen komplett zu verhindern. Zusätzlich können PETs bspw. die Offenlegung persönlicher Daten eines Nutzers durch einen anderen Nutzer nicht verhindern. Eine solche Datenpreisgabe durch Dritte kann u.a. im Rahmen der Synchronisierung des Adressbuches eines Smartphones mit einem Dienstanbieter wie Google oder Apple erfolgen.

Als Ergänzung zu PETs sollen deshalb Transparenz schaffende Mechanismen (*Transparency-Enhancing Technologies, TETs*) helfen, Nutzern die Möglichkeit zu geben, ihre persönlichen Daten zu schützen [Acc08]. Gemäß der Definition der Privatheit von Alan Westin [Wes67], an der sich auch die EU Richtlinie 95/46/EG zum Datenschutz orientiert [Eur95], beinhaltet das Recht auf Privatheit neben dem Recht des Individuums darauf zu entscheiden, wem gegenüber es welche persönliche Daten preisgegeben möchte, auch das Recht eines Individuums darauf, von Dritten über es gespeicherte Informationen einzusehen, zu ändern oder zu löschen. Um letzteres Recht ausüben zu können, müssen Nutzer datenzentrischer Dienste aber über Einblick in die über sie von Dienstanbietern gespeicherten Informationen verfügen. Der kombinierte Einsatz von Transparenzmechanismen, die genau dies ermöglichen sollen, und präventiven PETs verfügt über das Potential, es Nutzern datenzentrischer Dienste zu ermöglichen, ihre Privatheit zu schützen.

Eine Art von TETs sind sogenannte *Privacy Dashboards*. Ein Dashboard ist generell *„a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance“* [Few06]. Privacy Dashboards[1] sollen Nutzern nicht nur Einblick in die von einem Dienstanbieter über sie gespeicherten Daten gewähren, sondern auch die Möglichkeit bieten, gespeicherte Informationen zu ändern oder zu löschen. Viele Betreiber personalisierter Werbenetzwerke, wie bspw. Google[2] (siehe Abbildung 1), Yahoo![3] oder AOL[4], bieten ihren Nutzern bereits rudimentäre Formen solcher Dashboards an.

Allerdings weisen diese, von datenzentrischen Dienstanbietern selbst betriebenen, Dashboards aus Nutzersicht erhebliche Schwächen auf. Obgleich Dienstanbieter über ihre Dashboards völlige Transparenz über gespeicherte Daten eines Nutzers und deren Herkunft und Verwendung schaffen könnten, werden bislang nur sehr wenige dieser Informationen offengelegt. Wie in Abbildung 1 beispielhaft an *Googles Ad Preference Manager* dargestellt, werden in solchen Dashboards beispielsweise Informationen über abgeleitete, vermutete Nutzerinteressen angezeigt. Nutzer haben dabei jedoch keine Möglichkeit festzustellen, auf Grund welcher Informationen diese vermuteten Nutzerinteressen hergeleitet wurden und somit keine Möglichkeit, die Konsequenzen ihrer Interaktionen mit dem Dienstanbieter abzuschätzen. Zusätzlich besteht für Nutzer keine Möglichkeit, die

[1] Im folgenden nur als „Dashboards“ bezeichnet.
[2] http://www.google.com/settings/ads/onweb/
[3] http://info.yahoo.com/privacy/us/yahoo/opt_out/targeting/
[4] http://advertising.aol.com/advisibility

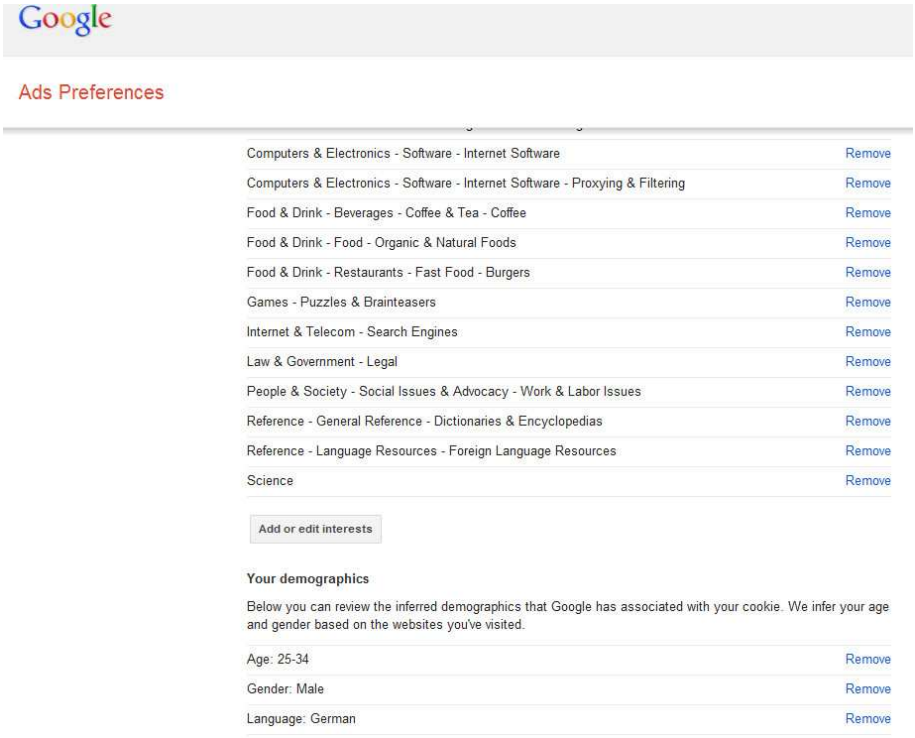

Abbildung 1: Googles *Ad Preference Manager* mit aus dem Nutzerverhalten abgeleiteten Werbekategorien

Korrektheit und Vollständigkeit der angezeigten Daten im Sinne einer Auditierung zu überprüfen. Das bedeutet, dass Nutzer blind darauf vertrauen müssen, dass der Dienstanbieter tatsächlich alle über sie gespeicherten Daten korrekt über das Dashboard anzeigt. Fehlende Anreize für datenzentrische Dienstanbieter, echte Transparenz herzustellen und gerechtfertigt mangelndes Vertrauen der Nutzer in momentane Dashboards verhindern, dass solche Dashboards ihr volles Potential als Mittel zur Realisierung einer holistischen Umsetzung von Privatheit ausschöpfen können.

Die Forschung im Bereich des Schutzes der Privatheit von Nutzern gegenüber Anbietern datenzentrischer Dienste ist hauptsächlich auf die Erforschung präventiver Datenschutztechniken fokussiert. Der in dieser Arbeit vorgestellte Ansatz zielt dagegen auf den Schutz der Privatheit mittels Transparenz und Kontrolle hinsichtlich bereits preisgegebener Daten ab. Diese Arbeit berichtet von unserer laufenden Arbeit daran, anbieterseitige Dashboards und die ihnen zugrunde liegende Infrastruktur so zu gestalten, dass sie einerseits Dienstanbietern Anreize für höhere Transparenz bieten können und andererseits Nutzern als vertrauenswürdiges Mittel und Nutzerschnittstelle zur Ausübung ihres Rechtes auf Privatheit dienen können. Unser Ansatz nutzt, als erster seiner Art, Methoden des Process Minings, um Transparenz mittels Dashboards vertrauenswürdig, d.h. auditierbar, zu er-

reichen. Diese Arbeit stellt die grundlegenden Komponenten unseres Ansatzes vor und diskutiert die generelle Umsetzbarkeit auditierbarer, vertrauenswürdiger Dashboards mittels der Verknüpfung von vorhandenen Methoden und Werkzeugen des Process Minings mit vertrauenwürdigen Logdateien auf Basis von Trusted Computing Platforms.

Folgender Abschnitt bietet einen kurzen Überblick über den Ansatz und stellt die dem Ansatz zugrunde liegenden Annahmen dar. In Abschnitt 3 stellen wir die einzelnen Komponenten des Ansatzes detaillierter vor. In Abschnitt 4 analysieren wir die Umsetzbarkeit unseres Ansatzes. Wir untersuchen den momentanen Stand der Forschung hinsichtlich Datenschutztechnologien und Dashboards in Abschnitt 5 und schließen unsere Arbeit in Abschnitt 6 mit einer Zusammenfassung und einem Ausblick auf noch ausstehende Forschung hinsichtlich unseres Ansatzes.

2 Grundlagen und Annahmen

Unser Ansatz hat das Ziel, Nutzern Transparenz und Kontrolle hinsichtlich von datenzentrischen Dienstanbietern über sie gespeicherter Daten und deren Verwendung zu gewähren. Ein Dashboard dient dabei als Nutzerschnitstelle. Die dem Dashboard zugrunde liegende Infrastruktrur dient dem Zweck, Logdateien des Dienstanbieters zu analysieren, um über Nutzer gespeicherte Daten zu identifizieren. Mittels Methoden des Process Minings soll dabei die Datenverwendung durch den Dienstanbieter erkannt werden.

Process Mining fokussiert, anders als klassisches Data Mining, nicht auf die Daten-, sondern auf die Prozessebene, um Einblicke aus prozessorientierter Sicht zu ermöglichen. Dazu sind die zu analysierenden Logdateien grundlegend aus Ereignissen aufgebaut, die Aktivitäten in Prozessen entsprechen. Ein Ziel des Process Minings ist es, Prozessmodelle aus Logdateien zu extrahieren [AUvdA12]. Der hier vorgestellte Ansatz nutzt Methoden des Process Minings zur Rekonstruktion von sequentiellen Abläufen[5] in den Systemen des Dienstanbieters.

Abbildung 2 zeigt einen schematischen Überblick über die Komponenten unseres Ansatzes. Wie in Abbildung 2 schematisch dargestellt, besteht der Ansatz aus fünf Komponenten. Um mittels Dashboards (5) nachvollziehbare und auditierbare Informationen bereitzustellen, werden die Logdateien (1) eines Dienstanbieters mit bekannten Verfahren des Process Minings (2) analysiert. Im Rahmen dieser Analyse werden Prozesse (3b) des Dienstanbieters und Datenverwendung durch den Anbieter erkannt und sogenannte *Log Views* (3a) erstellt. Die so gewonnenen Erkenntnisse können, nach entsprechender Aufbereitung (4), mittels Dashboards den Nutzern zugänglich gemacht werden. Eine detaillierte Beschreibung der einzelnen Komponenten erfolgt in Abschnitt 3.

Anbieterseitige Dashboards haben gegenüber nutzerseitigen Dashboards den Vorteil, dass nur über anbieterseitige Dashboards tatsächlich alle über einen Nutzer gespeicherten Daten angezeigt werden können. Nutzerseitige Dashboards können insbesondere nur bedingt Transparenz hinsichtlich der Datenverwendung durch einen Anbieter schaffen und

[5] Fortan als *Prozesse* bezeichnet

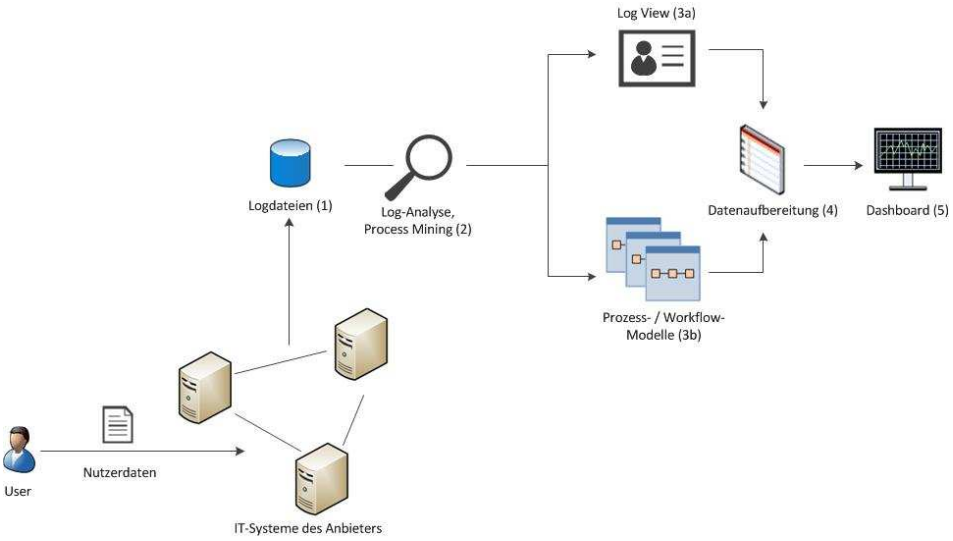

Abbildung 2: Informationsgewinnung für Dashboards mittels Process Mining

verfügen nicht über die Möglichkeit, Informationen anzuzeigen, die ein Anbieter aus anderen Quellen als dem Nutzer selbst gewonnen hat. Anbieterseitige Dashboards mittels Methoden des Process Minings aussagekräftiger und auditierbar zu gestalten, erfordert aber die Implementierung der dafür notwendigen Mechanismen auf Seiten des Dienstanbieters. Damit derart gestaltete Dashboards eine sinnvolle Ergänzung zu anderen TETs und nutzerseitig eingesetzten PETs darstellen können, müssen drei Voraussetzungen erfüllt sein auf die wir im Folgenden detailliert eingehen.

Annahme 1: Eine der Anwendungen des Process Minings ist die Rekonstruktion von Prozessmodellen aus Logdateien [AUvdA12, VdA11]. Unser Ansatz beruht auf der Annahme, dass Interaktionen von Nutzern mit den Systemen eines Dienstanbieters und die Reaktionen dieser Systeme darauf als Schritte von sequentiell ablaufenden Standardabläufen, d.h. Prozessen, dargestellt werden können.

Obgleich dies nicht in allen Fällen zutreffen muss, ist diese Annahme im Kontext datenzentrischer Webdienste haltbar. Einerseits folgen die meisten Nutzerinteraktionen mit den Diensten eines Anbieters festen Schemata. So besteht bspw. der Ablauf einer Suchanfrage eines Nutzers über eine Suchmaschine oder das Versenden einer Nachricht innerhalb eine Online Social Network Dienstes aus den immer gleichen Schritten. Auch wenn sich diese Schemata im Detail ändern können, führen Änderungen an diesen Abläufe nicht dazu, dass prinzipiell keinen Standardabläufen mehr gefolgt werden würde. Andererseits erfolgt die automatisierte Erstellung von Nutzerprofilen durch die Systeme eines Dienstanbieters ebenfalls anhand von Prozessen. Diese Annahmen werden durch von Dienstanbietern wie bspw. IBM [New06], Facebook [KCZ+09] oder Yahoo! [CKLY10] (siehe Abbildung 3) beantragten oder gehaltenen Patenten untermauert.

Annahme 2: Die Anwendung von Mechanismen des Process Minings erfordert die Exis-

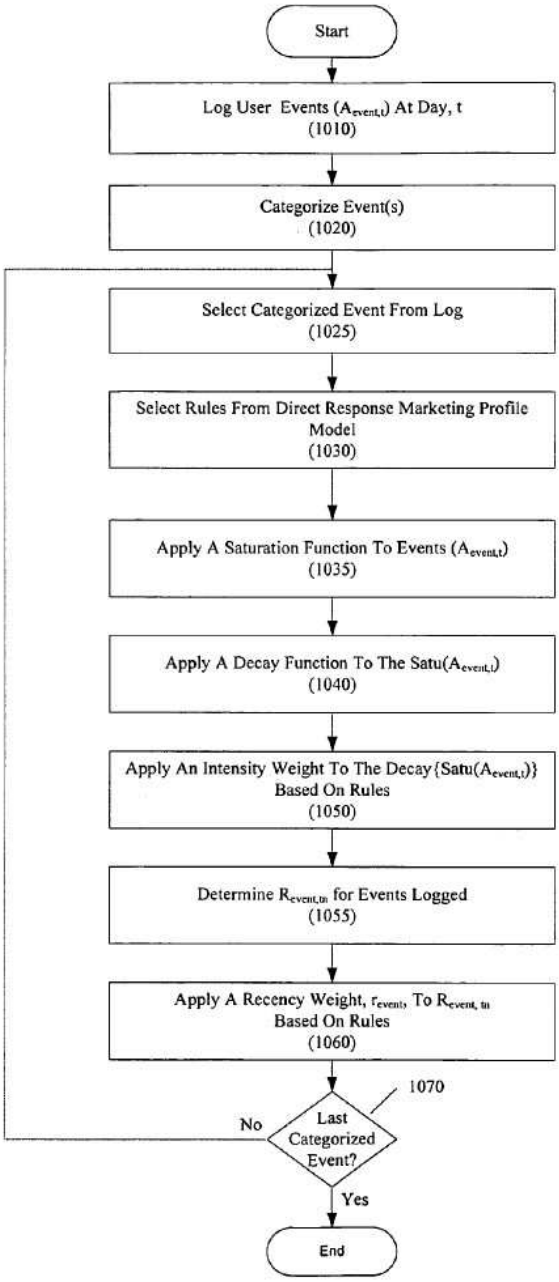

Abbildung 3: Ablauf der Berechnung eines Interesse-Maßes eines Nutzers durch Yahoo! [CKLY10]

tenz von Logdateien, auf denen diese Mechanismen angewendet werden können. In dieser Arbeit gehen wir von der Annahme aus, dass jedes Ereignis in den IT-Systemen eines Anbieters sowie jeder Zugriff auf gespeicherte Daten als Ereignis in Logdateien aufgezeichnet wird und die Logdateien so aufbereitet werden können, dass eine Analyse durch Methoden des Process Minings möglich ist. Zusätzlich gilt hier die Annahme, dass Logdateieinträge die die Daten eines spezifischen Nutzers betreffen dem Nutzer zugeordnet werden können.

Annahme 3: Das Anbieten von auf Process Mining beruhenden Dashboards verursacht einem Dienstanbieter Kosten. Zusätzlich besteht für einen Dienstanbieter das Riskio, dass ein zu aussagekräftiges Dashboard seinen Wettbewerbern unerwünschte Einblicke in die internen Prozesse des Anbieters ermöglichen könnte. Für einen Dienstanbieter ist das Anbieten eines unseres Ansatzes entsprechenden Dashboards daher nur vorteilhaft, falls der erwartete Nutzen der erhöhten Transparenz durch solche Dashboards diese Kosten und Risiken aufwiegen kann.

Aussagekräftige und nachweisbar vertrauenswürdige Dashboards verfügen aber über das Potential, Bedenken der Nutzer bezüglich des Schutzes ihrer Privatheit erheblich zu verringern. Dies wiederum kann zu verstärkter Nutzung der Dienste des Dienstanbieters führen und möglicherweise zu erhöhter Bereitschaft der Nutzer, persönliche Daten preiszugeben, was zu detaillierteren und somit wertvolleren Nutzerprofilen führen kann [CS05]. Eine detaillierte Untersuchung, welcher Grad an Transparenz durch Dashboards sowohl für Dienstanbieter wie auch für Nutzer vorteilhaft ist, steht bisher aus, überschreitet aber den Rahmen dieser Arbeit und erfolgt in zukünftiger Arbeit. In dieser Arbeit gehen wir vorläufig von der Annahme aus, dass vollständige Transparenz sowohl für Dienstanbieter wie auch ihre Nutzer vorteilhaft ist.

3 Komponenten des Ansatzes

Wie in Abbildung 2 schematisch dargestellt, besteht unser Ansatz aus fünf Komponenten. Im Folgenden stellen wir die einzelnen Komponenten unseres Ansatzes und den jeweiligen Stand ihrer Entwicklung vor.

3.1 Vertrauenswürdige Logdateien

Logdateien erlauben die Rekonstruktion vergangener Ereignisse eines IT-Systems. Im Falle prozessorientierter Systeme ermöglichen deren Logdateien auch die Rekonstruktion der abgelaufenen Prozesse. Somit kann rekonstruiert werden, auf Grund welcher Ereignisse der zum Rekonstruktionszeitpunkt vorgefundene Zustand des Systems erreicht wurde. Um allerdings Logdateien zum Zwecke eines Audits bzw. als vertrauenswürdigen Basis für in Dashboards angezeigte Informationen nutzen zu können, müssen diese Logdateien selbst korrekt und vertrauenswürdig sein, also die Eigenschaften der Integrität und Vertraulichkeit besitzen. Dies bedeutet, dass die Logeinträge korrekt (also die tatsächlichen Ereignisse

im System widerspiegeln) und vollständig sein (also alle Ereignisse des Systems widerspiegeln) müssen [Acc06]. Eine Beschreibung und eine prototypische Implementierung eines Systems zur Sicherstellung der Authentizität und Vertraulichkeit von Logdateien in verteilten Systemen wurde von Accorsi in [Acc13] vorgestellt. Die in [Acc13] vorgestellte *BBox* basiert auf Public Key Kryptographie und Trusted Computing Platforms und kann nach Vollendung der nötigen Anpassungen auch für Zwecke des in dieser Arbeit vorgestellten Ansatzes verwendet werden.

3.2 Logdateianalyse mittels Process Mining

Gemäß unserer, in Abschnitt 2 dargelegten, Annahmen können Nutzerinteraktionen mit den Diensten eines datenzentrischen Dienstanbieters sowie die Verarbeitung von Nutzerdaten durch die Systeme des Dienstanbieters als Prozesse dargestellt werden. Die in diesen Prozessen ausgeführten Aktivitäten spiegeln sich, gemäß unserer Annahmen, in den Logdateien des Dienstanbieters wider. Diese Logeinträge können dergestalt aufbereitet werden, dass die Erkennung von Prozessen durch die Verwendung bekannter Process Mining Methoden möglich ist. Erkannte Prozesse können genutzt werden, um Nutzern mittels eines Dashboards bspw. Auskunft über die Herkunft inferierter Daten oder über andere Verwendung von Nutzerdaten zu geben.

Beispiel: Im in Abbildung 4, stark vereinfacht, dargestellten Fall bekundet ein Facebook-Nutzer (UserId 12) sein Interesse an den Themen „Fussball“ (*topic 95*) und „Freiburg“ (*topic 30*) durch „liken“ (1) dieser Themen. Diese Nutzeraktionen hinterlassen in den Logdateien des Dienstanbieters Facebook entsprechende Spuren (2a). Wie bereits in Abbildung 3 am Beispiel Yahoo!s dargestellt, erfolgt auf eine Nutzerinteraktion mit einem Dienstanbieter die Ausführung eines auf die Interaktion reagierenden Prozesses, hier eines Kategorisierungsprozesses (2b). Der in Abbildung 4 schematisch dargestellte Kategorisierungsprozess ordnet aufgrund der von ihm verwendeten *Association Rules* den Nutzer in die Kategorie derer ein, die höchstwahrscheinlich auch am Thema „Sportclub Freiburg“ (*topic 4*) interessiert sind. Auch diese Kategorisierung hinterlässt, gemäß unserer in Abschnitt 2 getroffenen Annahmen, Spuren in den Logdateien.

Mittels der Rekonstruktion der abgelaufenen Nutzeraktionen und der Aktionen des Systems aus den Logdateien kann zusammen mit den im folgenden erläuterten Log Views einem Nutzer transparent gemacht werden, wieso er der Kategorie derer, die sich für das Thema „Sportclub Freiburg“ interessieren, zugeordnet wurde. Analog können anderen Aktionen des Systems eines Dienstanbieters auf den Daten eines Nutzers rekonstruiert werden, sofern sich diese Aktionen in den Logdateien widerspiegeln.

3.3 Log Views

Ähnliche dem namensgleichen Konzept aus dem Datenbankbereich, dienen Log Views dazu, Zugriff auf eine Untermenge der Einträge einer oder mehrerer Logdatei auf vereinfach-

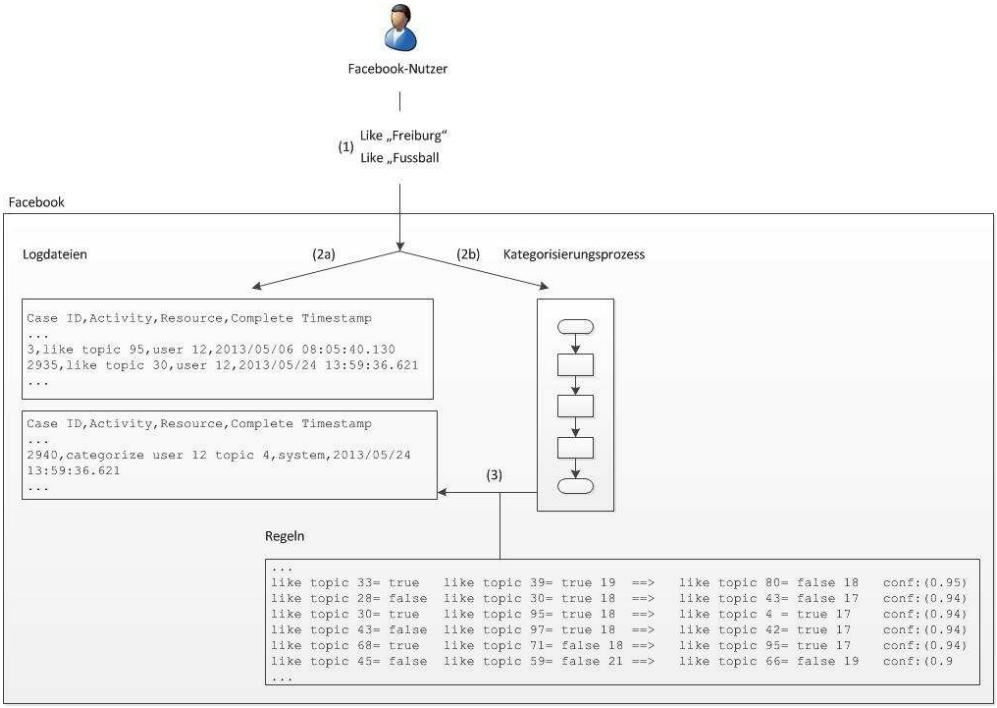

Abbildung 4: Vereinfachte Darstellung einer Nutzeraktion („like“) in Facebook

te Art zu ermöglichen. Im Kontext unserer Arbeit stellen Log Views eine nutzerspezifische Sicht auf die Logdateien eines Dienstanbieters dar, d.h. eine Auswahl aller Logeinträge, die einem spezifischen Nutzer zugeordnet werden können. Dies beinhaltet sowohl Logeinträge hinsichtlich Interaktionen des Nutzers mit dem Dienstanbieter, wie auch Logeinträge hinsichtlich der Verwendung der Daten des Nutzers durch den Dienstanbieter. Im Falle des in Abbildung 4 dargestellten Beispiels würde der Log View des Nutzers mit der UserID 12 sowohl die Logeinträge über die erfolgten „likes“ des Nutzers enthalten, wie auch den Eintrag über die erfolgte Zuordnung der Interessenkategorie *topic 4*. Das Konzept der Log Views wurden bereits 2006 von Sackmann et al. vorgestellt [SSA06].

Log Views stellen zusammen mit den durch die Prozessrekonstruktion erkannten Zusammenhängen die Grundlage für die Aufbereitung der Nutzerdaten zur Anzeige im Dashboard dar.

3.4 Datenaufbereitung und Privacy Dashboard

Um ihr Potential als Mittel zum Schutze der Privatheit der Nutzer datenzentrischer Dienste entfalten zu können, müssen Dashboards die vom Dienstanbieter gespeicherten Nutzerdaten und deren Verwendung durch den Dienstanbieter nicht nur überprüfbar korrekt und

vollständig, sondern auch übersichtlich und leicht verständlich anzeigen. Die benutzerfreundliche Gestaltung von Dashboards allgemein wird zwar bspw. in [Few06] behandelt. Unseres Wissens nach, ist aber die Frage, welche Informationen anbieterseitige Privacy Dashboards wie darstellen können, bisher nicht erforscht.

Verschiedene Ansätze zur nutzerfreundlichen Gestaltung von Dashboards sind denkbar. Nutzern könnte bspw. die Möglichkeit gegeben werden, bestimmte Typen von Daten und deren Verwendung auszuwählen, um diese prominent angezeigt zu bekommen. So könnte ein Nutzer etwa wählen können, per RSS-Feed über vom Dienstanbieter neu inferierte Informationen informiert zu werden und andere Typen von Daten nur über die Dashboard-Webseite zu überprüfen und ggf. zu korrigieren oder zu löschen. Eine solche Gestaltung von Dashboards würde eine Klassifizierung von Nutzerdaten in verschiedene Typen erfordern. Verschiedene Klassifizierungen von Nutzerdaten existieren bereits, bspw. von Schneier [Sch10] oder vom W3C [W3C02].

4 Analyse der Umsetzbarkeit des Ansatzes

Neben der Frage der technischen Umsetzbarkeit unseres Ansatzes, stellt sich auch die Frage, ob überprüfbar vollständige und korrekte Transparenz im Falle technischer Umsetzbarkeit im ökonomischen Sinne umsetzbar ist. Wie bereits in Abschnitt 2 erwähnt, besteht für Dienstanbieter nicht zwangsläufig ein Anreiz, vollständige Transparenz zu gewähren. Eine Untersuchung eines für Nutzer und Dienstanbieter vorteilhaften Grads an Transparenz steht bisher aber aus und überschreitet den Rahmen dieser Arbeit. Im Folgenden beschränken wir uns daher auf eine Untersuchung der Möglichkeiten der Umsetzung unseres Dashboard-Ansatzes in technischer Hinsicht.

Die Umsetzbarkeit unseres Ansatzes beruht auf der Umsetzbarkeit der einzelnen Komponenten des Ansatzes und der Möglichkeit die jeweiligen, teilweise bereits erforschten, Methoden und Mechanismen zu verknüpfen. Wie in Abschnitt 3 dargestellt, stellen vertrauenswürdige Logdateien die Grundlage unseres Dashboard-Ansatzes dar. Diese Logdateien müssen dabei nicht nur überprüfbar korrekt und vollständig sein, sondern auch Logeinträge dergestalt enthalten, dass die Rekonstruktion abgelaufener Prozesse und die Erstellung von Log Views möglich sind. Die in Abschnitt 3 bereits erwähnte *BBox* ermöglicht die Gewährleistung der Vertraulichkeit von Logeinträgen und ermöglicht die Erkennung von Manipulationen der Einträge [Acc13]. Die vollständige Implementierung und Anpassung von *BBox* an den Kontext anbieterseitiger Dashboards ist noch nicht beendet.

Um aus Logdateien Abläufe rekonstruieren und Log Views erstellen zu können, müssen die Logdateien entsprechend aufbereitet werden. Dazu müssen Ereignisse, die Aktivitäten in Prozessen entsprechen, in Logdateien identifiziert werden, um so sogenannte *Ereignislogs (Event Logs)* zu erstellen [AUvdA12]. Die Lösung des Problems, Ereignisse aus teilweise unstrukturierten und verteilt gespeicherten Logdateien zu extrahieren, um anhand dieser Ereignisse Abläufe rekonstruieren zu können, ist Hauptbestandteil jeder Prozess- oder Workflow-Rekonstruktion. Der Erfolg einer solchen Extraktion und die Qualität der so generierten Event Logs wird maßgeblich durch die Struktur der zugrunde liegenden

Logdateien bestimmt. Unter der Voraussetzung entsprechend umgesetzter Logging-Verfahren ist die Erstellung von Event Logs, auf denen Methoden des Process Minings angewendet werden können, zwar nicht-trivial aber dennoch prinzipiell umsetzbar.

Zahlreiche Methoden, um aus Event Logs abgelaufene Workflow- bzw. Prozessmodelle zu rekonstruieren, existieren, bspw. von van der Aalst et al. [VdAWM04] oder Cook und Wolf [CW98]. Implementierte Versionen diverser Workflow- bzw. Process-Mining-Algorithmen existieren und können in Frameworks wie *ProM* genutzt werden, um aus Event Logs Worksflow bzw. Prozess-Modelle zu rekonstruieren [vDdMV+05]. Eine Anwendung dieser Algorithmen auf aus Logdateien datenzentrischer Dienstanbieter extrahierten und entsprechend strukturierte Event Logs ist prinzipiell möglich.

Gemäß unserer in Abschnitt 2 dargelegten Annahmen, können Logeinträge, die die Daten eines Nutzers betreffen, diesem Nutzer zugeordnet werden. Obgleich dies in vielen Fällen zutrifft, hält diese Annahme nicht in allen Anwendungsfällen. Die Problematik der Erstellung von Log Views im Kontext von dynamischen Systemen wurde bereits in [SSA06] diskutiert. Vollständigkeit und Korrektheit von Log Views hängt direkt von der Struktur, Vollständigkeit und Korrektheit der zugrunde liegenden Logdateien ab. Unter der Voraussetzung entsprechend umgesetzter Logging-Verfahren ist die Erstellung vollständiger und korrekter Log Views, wie in [Acc08] anhand einer prototypischen Implementierung auf Basis von *BBox* gezeigt, möglich.

Dashboards als Nutzerschnittstellen zur Einsicht in die Ergebnisse der Analyse der Logdateien eines Anbieters müssen Nutzern auch die Möglichkeit bieten, über sie gespeicherte Daten zu verändern oder zu löschen. Die tatsächliche Durchführung einer Änderung oder Löschung muss dabei von Nutzern ebenfalls überprüfbar sein. Unter den von uns getroffenen Annahmen, würden sich solche Änderungs- bzw. Löschaktionen ebenfalls in den Logdateien des Dienstanbieters widerspiegeln und wären entsprechend überprüfbar. Eine nicht durchgeführte Änderung oder Löschung würde, unserer Annahmen nach, ebenfalls auffallen, sobald nur vordergründig geänderte oder gelöschte Informationen weiterhin in Logeinträgen von Aktionen des Systems des Anbieters erscheinen würden.

5 Stand der Forschung

Nutzer datenzentrischer Dienste sind bereit, begrenzte Mengen persönlicher Daten preiszugeben [AGDG06, GA05]. In Online Social Network Diensten beruht der Nutzen des Dienstes für die Nutzer sogar hauptsächlich auf der willentlichen Veröffentlichung persönlicher Daten durch die Nutzer [AGDG06]. Allerdings formulieren Nutzer ernste Bedenken hinsichtlich des Mangels an Transparenz und Kontrolle darüber, wie ihre Daten von datenspeichernden bzw. -verarbeitenden Stellen genutzt werden [PNF00].

Während PETs schon seit langer Zeit erforscht werden, sind TETs ein relativ junges Forschungsfeld. Eine Vielzahl von PETs, die automatisierte Nutzerprofilerstellung erschweren können, existiert bereits [VBBO03]. Auf Kryptographie basierende Ansätze wie etwa im, auf Onion Routing beruhenden, *TOR*-Projekt sollen die Anonymität eines Nutzers gewährleisten [DMS04]. Verschleierungstechniken, wie in *TrackMeNot* verwendet, haben

das Ziel, vor einem Dienstanbieter zu verbergen, für welche Themen sich ein Nutzer interessiert, um so die Kategorisierung des Nutzers zu erschweren [HN09]. Eine große Zahl weiterer präventiver Ansätze zum Schutze der Privatheit von Nutzern datenzentrischer Dienste existiert. Unseres Wissens nach existiert aber keine PET, die geeignet wäre, das Inferieren von Informationen über Nutzer aus bereits preisgegebenen Daten zu verhindern. Zusätzlich sind präventive Mechanismen des Datenschutzes konzeptionell prinzipiell nicht in der Lage, Nutzern Einsicht in bereits preisgegebene Daten zu gewähren.

Der Mangel an Kontrolle und Transparenz hinsichtlich bereits preisgegebener Daten und deren Verwendung hat in letzter Zeit zu verstärkter Forschung im Bereich der TETs geführt. Eine grundlegende Untersuchung und Klassifikation von TETs im Kontext von *Ambient Intelligence* führen Hildebrandt in [Hil09] und Hildebrandt et al. in [WP709] durch. Um Transparenz hinsichtlich geplanter Datennutzung durch einen Dienstanbieter herzustellen, wurde das Privacy Policy Format und Protokoll *P3P* entwickelt[6]. Nutzerseitige Dashboards wurden bspw. im *PrimeLife* Projekt untersucht und entwickelt [W3C11]. Das in *PrimeLife* entwickelte Dashboard ermöglicht es Nutzern, festzustellen, welche Cookies beim Besuch einer Webseite gespeichert werden und welche Daten übertragen werden. Diesem Dashboard mangelt es aber an der Möglichkeit, von Dienstanbietern aus anderen Quellen gewonnene Informationen anzuzeigen oder gar zu verändern.

Weitzner et al. stellen in [WABL$^+$06] den Ansatz des *Policy Aware Web* vor, der auf ähnlichen Prinzipien, wie der in dieser Arbeit vorgestellte Ansatz beruht. Anders als unser Ansatz zielen die in [WABL$^+$06] und [WABL$^+$08] vorgestellten Arbeiten auf die Überprüfung der Einhaltung von Datenverwendungspolicys ab und nicht auf die Offenlegung gespeicherter Nutzerdaten und die Rekonstruktion von, Nutzerdaten betreffenden, Prozessen.

Umfangreiche Forschung hinsichtlich sicherer Loggingverfahren existiert. Wie bereits in [Acc13] ausführlicher argumentiert, hat *BBox* gegenüber anderen Verfahren den Vorteil, dass Manipulationen der Logdateien erkennbar sind und die Möglichkeit der Schlüsselwortsuche in verschlüsselten Logdateien gegeben ist.

Banescu und Zanonne erforschen ebenfalls die Verwendung von Methoden des Process Minings zum Schutze der Privatheit. Auch Banescu und Zanonne untersuchen nicht Methoden, Verletzungen der Privatheit zu vermeiden, sondern Methoden, solche Verletzungen nachträglich zu erkennen. In [BZ11] stellen Banescu und Zannone einen Ansatz zur Erkennung von Verletzungen der Privatheit vor, der auf der Anwendung von Process Mining Methoden zur Erkennung von Diskrepanzen zwischen Prozessspezifikationen und tatsächlicher Prozessausführung beruht. In [PPZ11] stellen Petkovic et al. einen auf Process Mining Methoden beruhenden Ansatz der nachträglichen Datenverwendungskontrolle vor. Anders als der in diesem Artikel vorgestellte Ansatz, beruhen diese Ansätze zum Schutze der Privatheit aber nicht auf der Verknüpfung von Methoden der Prozessrekonstruktion mit Privacy Dashboards.

[6] http://www.w3.org/P3P/

6 Zusammenfassung und Ausblick

In dieser Arbeit haben wir unser Konzept, anbieterseitige Dashboards durch die Verwendung von Methoden des Process Minings auf Basis vertrauenswürdiger Logsdateien und Trusted Computing Platforms vertrauenswürdiger und aussagekräftiger zu gestalten, vorgestellt. Der vorgestellte Ansatz soll als Ergänzung zu PETs dienen, um Nutzern datenzentrischer Dienste Transparenz und Kontrolle hinsichtlich von Dienstanbietern über sie gespeicherter Daten und deren Verwendung zu gewähren.

Weiterer Forschungsbedarf besteht nicht nur hinsichtlich einer ökonomischen Analyse des vorgestellten Konzepts. Auch wenn wir in Abschnitt 4 die technische Umsetzbarkeit unseres Ansatzes gezeigt haben, steht eine komplette, prototypische Implementierung des Ansatzes noch aus. Einzelne Komponeten des Ansatzes sind bereits prototypisch implementiert, die Anpassung an den Kontext des Ansatzes und die Verknüpfung der Komponenten ist aber noch nicht abgeschlossen. Forschungsbedarf besteht auch hinsichtlich der Gestaltung von Privacy Dashboards als Nutzerschnittstelle. Um die Benutzerfreundlichkeit von Dashboards sicherzustellen ist des Weiteren eine Klassifikation persönlicher Nutzerdaten notwendig. Die Möglichkeit der Verknüpfung des in diesem Artikel vorgestellten Ansatzes mit Privacy-Policys bzw. Inference-Policys ist ebenfalls Gegenstand zukünftiger Forschung.

Literatur

[Acc06] R. Accorsi. On the Relationship of Privacy and Secure Remote Logging in Dynamic Systems. In S. Fischer-Hübner, K. Rannenberg, L. Yngström und S. Lindskog, Hrsg., *SEC*, Jgg. 201 of *IFIP*, Seiten 329–339. Springer, 2006.

[Acc08] R. Accorsi. *Automated counterexample-driven audits of authentic system records*. Dissertation, Albert-Ludwigs-Universität Freiburg, 2008.

[Acc13] R. Accorsi. A secure log architecture to support remote auditing. *Mathematical and Computer Modelling*, 57(7-8):1578–1591, 2013.

[AGDG06] A. Acquisti, Ralph Gross, G. Danezis und P. Golle. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In *PET*, Jgg. 4258, Seiten 36–58. Springer Berlin / Heidelberg, 2006.

[AUvdA12] R. Accorsi, M. Ullrich und W. van der Aalst. Process Mining. *Informatik Spektrum*, 35(5):354–359, 2012.

[AZM12] R. Accorsi, C. Zimmermann und G. Müller. On Taming the Inference Threat in Social Networks. In *1st International Workshop on Privacy and Data Protection Technology (PDPT)*, Amsterdam, 2012.

[BZ11] S. Banescu und N. Zannone. Measuring Privacy Compliance with Process Specifications. In *Security Measurements and Metrics (Metrisec), 2011 Third International Workshop on*, Seiten 41 –50, 2011.

[CKLY10] C. Chung, J. Koran, L. Lin und H. Yin. US Patent 7,809,740 B2, Model for generating user profiles in a behavioral targeting system, Oktober 2010.

[CS05] R. Chellappa und R. Sin. Personalization versus Privacy: An Empirical Examination of the Online Consumers Dilemma. *Information Technology and Management*, 6(2-3):181–202, April 2005.

[CW98] J. Cook und A. Wolf. Discovering models of software processes from event-based data. *ACM Trans. Softw. Eng. Methodol.*, 7(3):215-249, Juli 1998.

[DMS04] R. Dingledine, N. Mathewson und P. Syverson. Tor: the second-generation onion router. In *Proceedings of the 13th conference on USENIX Security Symposium - Volume 13*, SSYM'04, Seite 21-21, Berkeley, CA, USA, 2004. USENIX Association.

[Eur95] European Commission. Directive 95/46/EC of the European Parliament and of the Council of 24th October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities L281*, 38:31–50, 1995.

[Fac13] Facebook Reports Fourth Quarter and Full Year 2012 Results. [online], 2013. http://investor.fb.com/releasedetail.cfm?ReleaseID=736911, zuletzt aufgerufen am 07. Mai 2013.

[Few06] S. Few. *Information dashboard design*. O'Reilly, 2006.

[GA05] R. Gross und A. Acquisti. Information revelation and privacy in online social networks. In *ACM WPES*, Seiten 71–80. ACM, 2005.

[Goo13] Google's Income Statement Information. [online], 2013. http://investor.google.com/financial/tables.html, zuletzt aufgerufen am 07. Mai 2013.

[Hil09] M. Hildebrandt. Profiling and AmI. In K. Rannenberg, D. Royer und A. Deuker, Hrsg., *The Future of Identity in the Information Society*, Seiten 273–310. Springer, 2009.

[HN09] D. Howe und H. Nissenbaum. TrackMeNot: Resisting surveillance in web search. In I. Kerr, V. Steeves und C. Lucock, Hrsg., *Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society*, Seiten 417–436. Oxford University Press, Oxford, UK, 2009.

[KCZ+09] T. Kendall, M. Cohler, M. Zuckerberg, Y. Juan, R. Jin, J. Rosenstein, A. Bosworth, Y. Wong, A. D'Angelo und C. Palihapitiya. US Patent App. 12/193,702, Social Advertisements and Other Informational Messages on a Social Networking Website, and Advertising Model for Same, Juli 2009.

[MFP12] G. Müller, C. Flender und M. Peters. Vertrauensinfrastruktur und Privatheit als ökonomische Fragestellung. In J. Buchmann, Hrsg., *Internet Privacy*, acatech Studie, Seite 143–188. Springer Verlag, September 2012.

[MSLC01] M. McPherson, L. Smith-Lovin und J.M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, Seiten 415–444, 2001.

[New06] D. Newbold. US Patent 7,000,194, Method and system for profiling users based on their relationships with content topics, Februar 2006.

[PNF00] J. Phelps, G. Nowak und E. Ferrell. Privacy Concerns and Consumer Willingness to Provide Personal Information. *Journal of Public Policy & Marketing*, 19(1):27–41, 2000.

[PPZ11] M. Petkovic, D. Prandi und N. Zannone. Purpose Control: Did You Process the Data for the Intended Purpose? In W. Jonker und M. Petkovic, Hrsg., *Secure Data Management*, number 6933 in Lecture Notes in Computer Science, Seiten 145–168. Springer Berlin Heidelberg, Januar 2011.

[Sch10] B. Schneier. A Taxonomy of Social Networking Data. *IEEE Security & Privacy*, 8:88, 2010.

[Sol08] D. Solove. The End of Privacy? *Scientific American*, 299(3):100–106, September 2008.

[SSA06] S. Sackmann, J. Strüker und R. Accorsi. Personalization in privacy-aware highly dynamic systems. *Commun. ACM*, 49(9):32–38, 2006.

[VBBO03] G. Van Blarkom, J. Borking und J. Olk, Hrsg. *Handbook of Privacy and Privacy-Enhancing Technologies*. College bescherming persoonsgegevens, The Hague, The Netherlands, 2003.

[VdA11] W. Van der Aalst. *Process mining*. Springerverlag Berlin Heidelberg, 2011.

[VdAWM04] W. Van der Aalst, T. Weijters und L. Maruster. Workflow mining: discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142, 2004.

[vDdMV+05] B. van Dongen, A. de Medeiros, H. Verbeek, A. Weijters und W. van der Aalst. The ProM Framework: A New Era in Process Mining Tool Support. In G. Ciardo und P. Darondeau, Hrsg., *Applications and Theory of Petri Nets 2005*, number 3536 in Lecture Notes in Computer Science, Seiten 444–454. Springer Berlin Heidelberg, Januar 2005.

[W3C02] W3C. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. [online], 2002. http://www.w3.org/TR/P3P/, zuletzt aufgerufen am 09. Mai 2013.

[W3C11] W3C. Privacy Enhancing Browser Extensions. Technical report, W3C, 2011.

[WABL+06] D. Weitzner, H. Abelson, T. Berners-Lee, C. Hanson, J. Hendler, L. Kagal, D. McGuinness, G. Sussman und K. Waterman. Transparent Accountable Data Mining: New Strategies for Privacy Protection. Technical Report MIT-CSAIL-TR-2006-007, Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, Januar 2006.

[WABL+08] D. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler und G. Sussman. Information accountability. *Commun. ACM*, 51(6):82–87, Juni 2008

[Wes67] A. Westin. *Privacy and Freedon*. Atheneum, New York, 1967.

[WP709] WP7. Behavioural Biometric Profiling and Transparency Enhancing Tools. Bericht D7.12, April 2009.

[YLS+11] S. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng und Ho. Zha. Like like alike: Joint friendship and interest propagation in social networks. In *Proceedings of the 20th international conference on World wide web*, Seiten 537–546. ACM, 2011.