Kollaborative Identifikation von Datenschutzverstößen

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte
(bei Veröffentlichung)

Dissertation

von

Thorben Burghardt

aus Mannheim

Tag der mündlichen Prüfung:	10. November 2010
Erster Gutachter:	Prof. Dr.-Ing. Klemens Böhm
Zweiter Gutachter:	Prof. Dr. iur. Indra Spiecker genannt Döhmann LL.M.

Danksagung

Die Durchführung eines Dissertationsprojektes, wie ich jetzt weiß, funktioniert nicht im Alleingang. Allen sei Dank, die mich während dieses Projektes unterstützt haben.

Besonderer Dank gilt meinem Betreuer Prof. Dr. Klemens Böhm. Er hat mich in das wissenschaftliche Arbeiten und Schreiben eingeführt und stand jederzeit für Rückfragen zur Verfügung – an Weihnachten, tagsüber, nachts, immer. Danke auch für den respektvollen gegenseitigen Umgang. Weiter gilt mein Dank Dr. Erik Buchmann, der als zweiter Betreuer meiner Arbeit die Null-Versionen meiner Papiere lesen durfte – zu Beginn dieses Projektes manchmal eine Herausforderung. Der freundschaftliche Umgang zwischen uns hat zu einem sehr effizienten Arbeiten beigetragen.

Frau Prof. Spiecker danke ich für die Zweitbegutachtung meiner Arbeit und den ebenfalls unkomplizierten, effizienten Umgang. Prof. Christopher W. Clifton gilt mein Dank für die Zusammenarbeit, seine hilfreichen Anmerkungen und seinen Einsatz für meine Forschungsprojekte. Prof. Jürgen Kühling und seiner Arbeitsgruppe danke ich für die erfolgreiche Zusammenarbeit in unserem DFG-Projekt.

Ohne die Beiträge von Studenten, insbesondere bei der Implementierung und bei Nutzerstudien, wäre dieser Arbeit nicht möglich gewesen. Lev Povalahev hat maßgeblich zu der CSE-Studie, Jens Müller und Ursula Kotzur zu der LBS-Studie, Timo Wankmüller zu der Vollzugsdefizitanalyse und Martin Helfer zu CLEF beigetragen.

Dankt man normalerweise Freunden und Kollegen getrennt, so darf ich dies in Personalunion tun. Matthias, Clemens, Björn, Ando, Guido – wann immer ich eine inhaltliche Frage hatte, habt ihr mir geholfen; was immer die einschlägige Presse geschrieben hat, wir haben es diskutiert. Jonas, Ralf, Stephan, Kerstin – auch wenn wir uns über das Arbeiten gefunden haben und es manchmal nicht lassen können, darüber zu sprechen, so habe ich doch auch eine wunderbare Zeit mit Euch verbracht.

Meiner Familie sei besonderer Dank ausgesprochen. Vielleicht am meisten dafür, dass sie nie meinem Wunsch nachgekommen ist, mich nach neun Jahren von der Schule zu nehmen. Ihr habt mir immer den Rücken freigehalten.

Den zwei wichtigsten Menschen in meinem Leben danke ich zum Schluss. Ich danke Dir Sabine, dass Du mich nicht nur durch das Dissertationsprojekt begleitet, sondern auch währenddessen geheiratet hast. Dir Zoe danke ich, dass Du in dieser Zeit gesund und munter das Licht der Welt erblickt und meinem Leben neuen Wert gegeben hast. Ihr werdet immer Teil meines Herzens sein. Ich mache keine falschen Versprechungen – dass ich in Zukunft mehr Zeit haben werde, glauben wir ja alle nicht.

Karlsruhe, im November 2010 *Thorben Burghardt*

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Das Internet hat die Welt durchdrungen und bietet dadurch Unternehmen wie Privatpersonen nahezu uneingeschränkte Möglichkeiten des Informationsaustausches und der Kommunikation. Nutzten im Jahr 2000 noch etwa 400 Millionen Menschen das Internet, so sind es heute etwa 1,2 Milliarden [Poi10]. Das entspricht 20 Prozent der Weltbevölkerung. In Deutschland sind 70 Prozent der Bevölkerung online, bei den bis zu 30-Jährigen sogar deutlich über 90 Prozent [TNS09]. Auch an Ländergrenzen macht das Internet nicht halt. Theoretisch kann jeder Internetnutzer jede der über 100 Millionen Webseiten [Net06] auf über 625 Millionen Hosts [TNS09] besuchen. Die auf diesen Webseiten einem internationalen Publikum angebotenen Dienste sind vielfältig, wie etwa elektronische Behördengänge, Online-Einkäufe, Online-Enzyklopädien oder Online-Bankverkehr.

Die Verbreitung des Internets und dessen Nutzung als wesentliches Informations- und Kommunikationsmedium haben wiederum die Anforderungen an den Zugang zum Internet beeinflusst: Internetnutzer möchten zu jeder Zeit und über verschiedenste Medien detaillierte Informationen über alles und jeden abrufen. So erfolgt der Zugriff auf diese Dienste nicht mehr nur über den fest installierten Rechner im Haushalt. Vielmehr durchdringen mobile Endgeräte den Alltag. In Deutschland hat das ‘mobile Internet’ bereits heute mehr als fünf Millionen Nutzer – mit stark steigender Tendenz [Deu10].

Die Entwicklung hin zur Nutzung von Online-Diensten wird dabei nicht nur durch den Komfort für den Nutzer vorangetrieben. Die Unternehmen sind sich der Allgegenwart des Internets bewusst und setzen ihren Nutzern Anreize, ihr jeweiliges Online-Angebot einzusetzen [And09]. So kostet eine Überweisung offline meistens Geld, online ist sie in der Regel kostenlos, offline kostet das Versenden von Nachrichten Geld, online ist dies kostenlos, offline kosten Zeitungen Geld, online sind viele kostenlos. Alleine schon diese monetären Anreize führen dazu, dass fast jeder Deutsche irgendwelche Online-Dienste nutzt. Nicht zuletzt sind Online-Informationen oftmals aktueller als beispielsweise Offline-Zeitungen, vielfach sind Online-Informationen sogar live.

Vor der Verwendung eines Online-Dienstes muss ein Nutzer sich im Allgemeinen registrieren – durch die Angabe personenbezogener Daten. Bei der Nutzung eines Dienstes geben Personen bewusst wie unbewusst weitere, unter Umständen sensible Daten preis. Bei einer Milliarde Menschen, die zusammen mehrere Millionen Online-Dienste nutzen, bedeutet dies eine riesige Menge personenbezogener Daten – eine Bedrohung für die Privatheit jedes einzelnen Internetnutzers und der Kontext dieser Arbeit.

1.1. Veränderung des Internets mit dem Web 2.0

Das Web 1.0 steht für Kommunikations- und Handelsbeziehungen zwischen Unternehmen und Privatpersonen (B2C) [Kno03]. Das Web 2.0 hat eine neue, grundlegende Art der Interaktion im Internet etabliert. Auch wenn es, insbesondere durch die Vermischung klassischer Web 1.0- und Web 2.0-Dienste, keine scharfe Abgrenzung des Begriffs Web 2.0 gibt, so zeichnet es sich doch durch zwei zentrale Merkmale aus [O'R05]: Anstelle weniger Anbieter, die für die breite Masse Dienste im Internet anbieten, stellen Anbieter im Web 2.0 vielmehr *Plattformen* bereit. Nutzer sind auf diesen Plattformen sowohl Inhaltsproduzent als auch Inhaltskonsument. Außerdem *kollaborieren* die Nutzer, das heißt, sie wirken bei der inhaltlichen Ausgestaltung der Dienste zusammen. Das Paradebeispiel für solch eine Kollaboration ist Wikipedia[1]. Hier tragen Millionen von internationalen Nutzern Inhalte zu einer Online-Enzyklopädie zusammen. Die Nutzer profitieren auf der einen Seite von Artikeln anderer Autoren, während sie als Autor selbst Informationen bereitstellen. Vielfach stellen die Plattformen Programmierschnittstellen (engl. application programming interface, API) bereit, um Entwicklern die Erstellung von *Mashups* zu ermöglichen. Mashup bezeichnet die Neukombination bestehender Inhalte, zum Beispiel Online-Karten mit Flugdaten zur Veranschaulichung von Flugbewegungen[2]. Das heißt, im Web 2.0 kann auf den gleichen Daten eine Vielzahl unterschiedlicher Dienste aufbauen.

Auch die Inhalte, die im Web 2.0 verglichen mit dem Web 1.0 ausgetauscht werden, haben sich verändert. Inhalte entstehen von Nutzern für Nutzer in einer Mischung aus Kommunikation und Information. So spricht man auch vom 'sozialen Web' als Teilgebiet des Web 2.0 [EGH08]. Die populärsten und meist frequentierten Anwendungen für solch einen sozialen, gegenseitigen Austausch sind sicherlich soziale Netzwerkseiten. Beispiele für Deutschland sind die VZ-Netzwerke, Stayfriends, Wer-Kennt-Wen oder Xing. Weltweit bieten MySpace, LinkedIn und mit 400 Millionen Nutzern Facebook als größte soziale Netzwerkseite vergleichbare Plattformen an [sti10]. Neben einem persönlichen Profil, über das sich die Nutzer darstellen können, bieten diese Plattformen Kommunikationsmöglichkeiten zwischen den Nutzern an, automatisch generierte Mitteilungen über Aktualisierungen in Profilen von Freunden, die Organisation in Interessensgemeinschaften, die Einbindung von Applikationen von Drittanbietern und vieles mehr.

Auch im Web 2.0 spielen mobile Endgeräte eine immer wichtigere Rolle. Mobile Endgeräte erlauben es Anbietern, ihre Dienste am Kontext der Nutzer auszurichten. Der Kontext ist dabei insbesondere der Standort einer Person. Dieser kann über erreichbare Drahtlosnetzwerke (WLAN) im Umfeld des mobilen Gerätes [RMT+02], die Mobilfunkwabe, mit der ein Endgerät verbunden ist, oder über exakte Geo-Koordinaten

[1] `http://wikipedia.org/`, Juni 2010
[2] `http://radar.zhaw.ch/`, Juni 2010

(Global Positioning System, GPS) [RM03]) ermittelt werden. Kollaborativ können sich Nutzer abhängig von ihrem Standort über Staus oder Verkehrskontrollen wie bei FoxyTag[3] benachrichtigen, können sich über Aufenthaltsorte von Freunden informieren [Tim08] oder bei der gemeinschaftlichen Annotation von Orten mit Fotos, wie bei dem Dienst Panoramio[4], mitwirken.

1.2. Privatheitsprobleme durch das Web 2.0

Daten, die Nutzer im Web 2.0 preisgeben, können von anderen Nutzern und den Diensteanbietern zur Erstellung umfangreicher Profile über Interessen, Gewohnheiten, Besitzverhältnisse etc. herangezogen werden. Diese können dann beispielsweise zur Bewertung des Gesundheitszustandes oder der Arbeitsfähigkeit einer Person [Ser10], der Bewertung der Kreditwürdigkeit oder der Einordnung einer Person in einen Versicherungstarif [Odl03] als auch für personalisierte Werbung [LSY03] und Preisdiskriminierung [Odl03, BM02] verwendet werden. Die preisgegebenen Informationen können aber auch zu Bedrohungen wie Cyber-Mobbing [LHU09] führen. Die Liste der Beispiele lässt sich beliebig fortsetzen.

Relevant für diese Arbeit ist eine Unterteilung dieser Probleme in zwei Klassen. Die Unterteilung entspricht den beiden genannten zentralen Prinzipien des Web 2.0, nämlich das Prinzip, das Web in Form von Anbietern von Plattformen zu organisieren und dem Prinzip der Kollaboration von Nutzern. Die Klassen sind (1) Privatheitsprobleme von Nutzern untereinander und (2) zwischen Nutzern und Anbietern.

Privatheitsprobleme Nutzer-Nutzer Durch die Doppelfunktion des Nutzers als Inhaltsproduzent und Inhaltskonsument fließen Inhalte zwischen den Nutzern. Das sind insbesondere im sozialen Web oftmals sensible, personenbezogene Informationen.

Privatheitsprobleme Nutzer-Anbieter Grundsätzlich bedarf es eines Betreibers für jede Web 2.0-Plattform, zum Beispiel Facebook als Anbieter einer sozialen Netzwerkseite. Dieser stellt die Infrastruktur, über die die Nutzer Inhalte bereitstellen und konsumieren, zur Verfügung. Im Regelfall bedeutet dies, dass der Plattformbetreiber dadurch sowohl Zugriff auf alle über die Plattform bereitgestellten Inhalte hat, als auch umfassende Einblicke, wer welche Inhalte konsumiert.

Für beide Klassen von Privatheitsproblemen gilt, dass die Bedrohung für das Individuum durch die Verknüpfung unterschiedlicher Informationsquellen verstärkt wird. So können ein Anbieter aber auch andere Nutzer aus dem Bewegungsprofil eines standortbezogenen Dienstes unter Umständen die Wohnadresse einer Person ermitteln und über die Adresse mit Angaben einer sozialen Netzwerkseite korrelieren. Die meisten Nutzer sind sich nicht bewusst, dass zu dem Bewegungsprofil dadurch auch Freunde, Interes-

[3] http://www.michelderiaz.com, Mai 2010
[4] http://www.panoramio.com, Mai 2010

sen und Gewohnheiten der betroffenen Person sichtbar sind. Gerade standortbezogene Dienste sind außerdem ein gutes Beispiel für die automatisierte Erhebung und Verarbeitung von Daten, das heißt im Hintergrund und vor dem Nutzer verborgen. Dass die daraus resultierenden Bedrohungen für die Privatheit nicht theoretischer Natur sind, zeigt eine Vielzahl alarmierender Pressemeldungen [BTZ06, Kan06, The07, HR06, Una09], Statistiken zu Verstößen, wie die von 'Privacy Right Clearinghouse'[5] oder dem 'Identity Theft Ressource Center'[6], sowie Forschungsarbeiten [Gol06, Swe00, PCT06, GA05]. Das Persistentmachen einmal preisgegebener sensibler Daten, führt zu einer unwiderruflichen Verletzung der Privatheit der betroffenen Person [Ros07].

1.3. Allgemeine Ansätze zum Schutz der Privatheit

Privatheitsproblemen kann auf unterschiedliche Weise entgegengetreten werden. Die drei zentralen Ansätze [WLW98, SGB01] sind (1) Technologien zum Schutz der Privatheit, (2) die Datenschutzgesetzgebung und (3) die Selbstregulierung von Unternehmen [Hoo05].

Selbstregulierung bedeutet, dass sich Unternehmen, zum Beispiel einer Branche, zu einem ethisch korrekten Umgang mit personenbezogenen Daten verpflichten. Von Interesse für diese Arbeit sind Technologien zum Schutz der Privatheit (1) unter Berücksichtigung der Datenschutzgesetzgebung (2).

Wir unterscheiden, insbesondere in Hinblick auf (2), die Begriffe Privatheit und Datenschutz. Beide haben ihren Ursprung in der Definition von Westin 'Privatheit ist der Anspruch von Personen oder Gruppen, selbst darüber zu verfügen, wann, wie und in welchem Umfang persönliche Daten an andere weitergegeben werden' [Wes67]. Wir werden den Begriff Privatheit, angelehnt an den Begriff 'Privacy' aus dem Englischen, als Oberbegriff benutzen. Das beinhaltet insbesondere all das, was Personen subjektiv als schützenswert, das heißt als privat, empfinden. Betrachten wir den Informationsfluss und die damit einhergehenden Probleme zwischen Nutzern, so sprechen wir von Privatheit. Betrachten wir das Nutzer-Anbieter Verhältnis, so interessiert uns in dieser Arbeit die datenschutzrechtliche Situation, wir sprechen also von Datenschutz und Datenschutzverstößen. Um für diese Arbeit einen handhabbaren Rechtsrahmen aufzuspannen, konzentrieren wir uns auf Dienste, die unter die Gesetzgebung gemäß des Bundesdatenschutzgesetzes (BDSG) fallen, insbesondere den Teil für den nichtöffentlichen Bereich. Konkreter ausgedrückt sind alle im Folgenden als (Dienste-) Anbieter bezeichneten Unternehmen Telemedienanbieter im Sinne des Telemediengesetzes (TMG). Trotz dieser Einschränkung umfasst das die meisten populären Anbieter wie soziale Netzwerkseiten, Foren, Shops etc.

Die Idee des Privatheitsschutzes durch Technik (3) ist nicht neu. Sie entstammt der

[5] `http://www.privacyrights.org/ar/ChronDataBreaches.htm`, Mai 2010

[6] Breach Database, `http://www.idtheftcenter.org`, Mai 2010

Forderung nach einem Systemdatenschutz, bei dem, durch den geeigneten Einsatz von Technik und organisatorischen Vorkehrungen für den vertrauensvollen Umgang mit personenbezogenen Daten, die Privatheit der Nutzer geschützt werden soll[7]. Die zum Schutz der Privatheit eingesetzte Technik wird als 'Privacy-Enhancing Technology' (PET) bezeichnet. Diese kann sehr einfach geartet sein, beispielsweise ein Häkchen, das ein Datum als privat definiert [LB08]. Sie kann aber auch sehr komplex sein, wie bei der Anonymisierung von Bewegungsprofilen [MCA06, GL04, KYS05, GG03] oder Suchanfragen [Ada07, KNPT07, KKMN09, MC08b].

1.4. Problemstellung dieser Arbeit

Wie beschrieben existieren technische Mechanismen zum Schutz der Privatheit. Trotzdem gibt es fast täglich Berichte über Privatheits- und Datenschutzprobleme. Wir untersuchen diesen Widerspruch. Dabei stellen sich folgende Probleme:

Es ist unklar, welche Privatheitspräferenzen Nutzer im Web 2.0 zum Schutz ihrer Daten haben. Ebenso ist unklar, wie ein Mechanismus ausgestaltet sein muss, damit ein Nutzer mit diesem seine Privatheitspräferenzen ausdrücken kann. Mechanismen unterscheiden sich aber auch in der Nutzungskomplexität, dem Nutzungsaufwand, dem erforderlichen technologischen Verständnis, dem Grad des erforderlichen Bewusstseins des Nutzers etc. Ein Nutzer muss seine Präferenz mit einem Mechanismus nicht nur ausdrücken, sondern dem Mechanismus auch vertrauen und ihn richtig anwenden können. Es ist eine offene Frage, welche Anforderungen an PETs das nach sich zieht.

Privatheitsprobleme entstehen nicht nur durch das Nutzerverhalten. Der Gesetzgeber definiert im Datenschutzgesetz Anforderungen an die Diensteanbieter, die einen vertrauensvollen Umgang mit personenbezogenen Daten sicherstellen sollen. Es ist aktuell unklar, inwieweit sich die Anbieter an diese Vorgaben halten. Da Internetnutzer im Allgemeinen keine Datenschutzexperten sind, ist außerdem fraglich, ob und wie Nutzer Datenschutzverstöße identifizieren können.

Personenbezogene Daten werden nicht zum Selbstzweck erhoben, sondern aufgrund funktionaler Anforderungen oder ökonomischer Interessen. Es ist eine weitere offene Frage, ob es einen Mittelweg gibt, so dass die Privatheit der Nutzer geschützt wird, gleichzeitig jedoch die funktionalen Anforderungen und ökonomischen Interessen berücksichtigt werden.

[7] https://www.datenschutzzentrum.de/systemdatenschutz/, Juni 2010

1.5. Ziele dieser Arbeit

Es ist das Ziel dieser Arbeit, anhand von Nutzerstudien herauszufinden, welche Privatheitspräferenzen Anwender von Web 2.0-Diensten haben. Dabei fokussieren wir insbesondere auf Präferenzen, die Nutzer zum Schutz vor anderen Nutzern definieren. Weiter möchten wir Anforderungen herausarbeiten, wie Mechanismen zum Schutz der Privatheit im Web 2.0 ausgestaltet sein müssen, (i) um dem Schutzbedürfnis der Nutzer zu genügen und um (ii) vom Nutzer akzeptiert und richtig eingesetzt zu werden. Nur wenn (i) und (ii) erfüllt sind, können Nutzer ihre Präferenzen auch tatsächlich durchsetzen. Es ist ebenfalls unser Ziel, diese Erkenntnisse an konkreten Implementierungen von neuartigen Web 2.0-Diensten zu gewinnen, die kurz vor der Durchdringung des Alltages stehen. So können unsere Ergebnisse den Entwicklern dieser und verwandter Dienste helfen, künftig effektive PETs zu entwerfen, das heißt, noch bevor eine Vielzahl von Nutzern private Informationen für immer preisgegeben hat.

Weiter gilt es zu untersuchen, inwieweit sich die Plattform- und Diensteanbieter im Web 2.0 konform zu geltendem Datenschutzrecht verhalten. Wir wollen herausfinden, inwieweit ein Nutzer, der bewusst mit seinen personenbezogenen Daten umgeht und PETs korrekt einsetzt, davon ausgehen muss, dass seine Bemühungen vergebens sind, da der Anbieter widerrechtliche Daten erhebt, speichert, nutzt oder weitergibt. Auch hier ist es wichtig zu erkennen, warum existierende PETs nicht greifen. Darauf aufbauend ist es unser Ziel, selbst ein PET zur systematischen Identifikation von Datenschutzverstößen zu entwerfen. Insbesondere soll Nutzern ohne datenschutzrechtliche Vorbildung die Identifikation von Verstößen ermöglicht werden.

Zuletzt ist es unser Ziel, einen Kompromiss für Nutzer und Anbieter zu finden. Dieser soll die Privatheit der Nutzer schützen und gleichzeitig Problemen der Anbieter beim Umgang mit personenbezogenen Daten entgegenwirken. Wichtig dabei ist, dass der Nutzen, den Anbieter aus dem Einsatz personenbezogener Daten und Profile ziehen, nicht verloren geht.

Das Erreichen dieser Ziele ist schwierig. Die Dienste, anhand derer wir die Privatheitspräferenzen identifizieren und PETs testen möchten, sind neu und existieren nur in Form von Prototypen. Wir müssen die Dienste also selbst und voll funktionstüchtig implementieren. Das erfordert fundierte Kenntnisse der eingesetzten Technologie und kostenintensive Hardware, auf der die Anwendungen laufen. Solche Hardware umfasst für standortbezogene Dienste beispielsweise neuste Mobiltelefone für alle Teilnehmer einer Studie. Außerdem sind die Dienste auch für die Nutzer neu. Für Nutzerstudien mit relevanten Ergebnissen müssen wir die Teilnehmer intensiv auf den neuen Technologien trainieren. Auch die Untersuchung der Diensteanbieter auf Datenschutzverstöße ist schwierig. Aktuell gibt es kein standardisiertes Vorgehen zur Verstoßidentifikation und das erst recht nicht für die Vielzahl unterschiedlicher Dienste im Internet. Eine technische Unterstützung fehlt für die meisten Verstöße ebenfalls, die Identifikation muss

somit zumeist händisch erfolgen. Soll ein PET die Identifikation von Verstößen auch Personen ohne datenschutzrechtliche Kenntnisse ermöglichen, muss nicht nur das PET entwickelt, sondern dieses ebenfalls wieder anhand von Nutzerstudien evaluiert werden. Zum Finden eines Kompromisses zwischen Privatheit für die Nutzer und Nutzen für die Anbieter ist es erforderlich, Daten zu erheben und einzusetzen, die eine realistische Privatheitsbedrohung darstellen und einen realistischen (ökonomischen) Nutzen widerspiegeln.

1.6. Beiträge dieser Arbeit

Die vorliegende Arbeit ist in vier Beiträge unterteilt. Zuerst betrachten wir im Zuge eigener Vorarbeiten das Nutzer- und Anbieterverhalten bei der Dienstnutzung. Die dabei gewonnenen Erkenntnisse vertiefen wir in den mittleren beiden Beiträgen, die wir gemäß der eingeführten zwei Klassen von Privatheitsproblemen in den 'Schutz der Privatheit zwischen Nutzern' und den 'Schutz der Privatheit vor den Anbietern' unterscheiden. Zuletzt untersuchen wir einen Ansatz, der beiden Klassen von Privatheitsproblemen gleichzeitig entgegenwirkt.

1.6.1. Nutzer- und Anbieterverhalten bei der Dienstnutzung

Der erste Teil der Arbeit umfasst eigene Vorarbeiten. Wir erheben hier in der Breite den Status Quo von sowohl dem Umgang von Nutzern mit ihren personenbezogenen Daten, als auch den Umgang der Anbieter mit diesen Daten. Diese Vorarbeit ist insbesondere durch das 2007 in Kraft getretene Telemediengesetz erforderlich geworden. Mit diesem Gesetz hat sich der Rechtsrahmen für Anbieter und Nutzer und somit auch der Ausgangspunkt für die Forschung im Kontext Datenschutz verändert.

Wir haben eine Nutzerstudie durchgeführt, Anbieter händisch analysiert und das Anbieterverhalten bei einer E-Mailanfrage untersucht.

Unsere Ergebnisse zeigen, dass fast alle Nutzer mit der Kontrolle ihrer personenbezogenen Daten überfordert sind. Nutzer wissen nicht, wo sie registriert sind, können sich nicht vorstellen, anhand welcher Daten sie identifiziert werden können und geben Daten versehentlich preis. Unternehmen erschweren die Kontrolle maßgeblich, indem sie nicht ausweisen, welche Daten sie zu welchem Zweck erheben, Empfänger personenbezogener Daten in unspezifischen Formulierungen verschleiern und Anfragen nur unzureichend beantworten.

Die folgenden Beiträge untersuchen diese Aspekte im Detail.

1.6.2. PETs im Web 2.0 zum Schutz zwischen Nutzern

Unser Ziel im zweiten Teil der Arbeit ist es herauszufinden, was Nutzer als schützenswert empfinden, was für PETs sie anwenden und wie sie sich beim Umgang mit un-

terschiedlichen PETs verhalten. Dazu haben wir speziell Web 2.0-Dienste untersucht, bei denen Nutzer kollaborieren und die dabei generierten Inhalte untereinander austauschen.

Wir haben unsere Fragestellungen anhand von PETs, die wir in zwei konkrete, von uns entwickelte Anwendungen integriert haben, evaluiert. Bei den Anwendungen handelt es sich zum einen um kollaborative Suchmaschinen (engl. Collaborative Search Engine, CSE). Bei diesen unterstützen sich Nutzer gegenseitig beim Auffinden von Informationen im Web, zum Beispiel durch den Austausch von Suchanfragen oder von relevanten Seiten zu einem Informationsbedürfnis. Suchanfragen können private Information beinhalten, die Rückschlüsse auf die suchende Person, deren Interessen und Gewohnheiten zulassen [BTZ06, Kan06, Dat08]. Zum Anderen betrachten wir repräsentativ für standortbezogene Dienste eine von uns entwickelte Anwendung, bei der Nutzer Orte abhängig von ihrer aktuellen Position über ein mobiles Endgerät mit Schlagworten annotieren. Die Schlagworte helfen anderen Nutzern beim Auffinden interessanter Orte. Die Kombination aus Schlagwort, Ort und Zeitpunkt der Erstellung kann jedoch eine sensible Information sein.

In zwei Studien zeigen wir, dass wir die Privatheitspräferenzen von Nutzern auf einfach strukturierte Regeln abbilden können. Für die Personengruppen, mit denen die Nutzer intensiv Inhalte austauschen, definieren Nutzer feingranular, wer, wann, welche Informationen sehen darf. Dies unterstreicht den Bedarf flexibler PETs. Zwei Arten von Strategien, die wir beobachtet haben, sind hervorzuheben: Erstens definieren Nutzer reziproke Strategien, bei denen sie dann Informationen teilen, wenn die potentiellen Empfänger dies auch tun würden. Zweitens fordern Nutzer Anonymität. Das heißt, kann eine Information einem Nutzer nicht zugeordnet werden, ist er bereit auch eigentlich sensible Daten zu teilen. Weiter haben wir gezeigt, dass einfache Mechanismen, die jedoch kontinuierlich das Bewusstsein des Anwenders fordern, versagen. Interessant dabei ist, dass Nutzer diese Unzulänglichkeit wahrnehmen. So wünschen sie die Kombination einfacher PETs, die sie leicht verstehen und nutzen können, mit komplexen, schwer verständlichen aber automatisiert arbeitenden Mechanismen.

1.6.3. Identifikation von Datenschutzverstößen der Diensteanbieter

Der dritte Teil der Arbeit konzentriert sich auf Datenschutzpraktiken von Web 2.0-Diensteanbietern und den Schutz der Privatheit der Nutzer vor dem Anbieter. Wir möchten herauszufinden, inwieweit ein Nutzer, der bewusst mit seinen personenbezogenen Daten umgeht und auch PETs korrekt einsetzt, durch die Datenschutzpraktiken der Anbieter gefährdet wird.

Die Identifikation bei den Anbietern vorherrschender Defizite ist zentral, damit Mechanismen zum Schutz von Nutzern untereinander überhaupt wirken können. Am Beispiel: Was hilft es, dass Nutzer die Sichtbarkeit ihrer Daten für andere Nutzer einschränken können, wenn der Anbieter den gesamten Datenbestand publiziert [BTZ06]

oder verkauft [Aig10]?

Vollzugsdefizitanalyse Eine mögliche Ursache für die Vielzahl der Datenschutzverstöße bei Online-Diensten sehen Datenschutzexperten in einem Vollzugsdefizit des Datenschutzrechts [WLW98]. Das heißt, Gesetze existieren, die Aufsichtsbehörden kontrollieren deren Einhaltung und Durchsetzung jedoch nicht ausreichend. Somit besteht für die Unternehmen wenig Anreiz, geltendes Recht einzuhalten.

Um das Ausmaß des angenommenen Vollzugsdefizits zu untersuchen, haben wir in einer interdisziplinären Studie mit Juristen der Arbeitsgruppe von Prof. Kühling (Universität Regensburg) einen Katalog von möglichen Datenschutzverstößen bei Online-Diensten entwickelt. Anhand 100 ausgewählter Anbieter haben wir die von extern erkennbare Datenschutzpraktik der Anbieter mit dem Verstoßkatalog abgeglichen – mit alarmierendem Ergebnis. Nur fünf der untersuchten Anbieter verhalten sich konform zu geltendem Recht. Insgesamt verstoßen die 100 Anbieter mehr als 300-mal gegen das Datenschutzrecht. Damit belegen wir die Vermutung des Vollzugsdefizits.

Kollaborative Identifikation von Datenschutzverstößen Weiter zeigen unsere Ergebnisse der Vollzugsdefizitanalyse, dass sowohl den Nutzern als auch den Anbietern die Kompetenz fehlt, selbst Datenschutzverstöße zu identifizieren. Den Aufsichtsbehörden wiederum fehlen die Ressourcen für eine effiziente Kontrolle der Anbieter. In Schleswig-Holstein[8] sind beispielsweise nur sechs Angestellte der Aufsichtsbehörde für die Internetauftritte von mehr als 100.000 Unternehmen verantwortlich. Eine effektive Überwachung ist somit unmöglich – mit dem im vorangegangenen Abschnitt beschriebenen Ergebnis.

Wir haben auf diesen Erkenntnissen aufbauend einen Ansatz zur kollaborativen Identifikation von Datenschutzverstößen entwickelt. Dazu haben wir juristische Expertise zur Identifikation von Datenschutzverstößen in Software gegossen. Die kollaborative Komponente unseres Ansatzes erlaubt es Nutzern, einen Beitrag zur Identifikation von Verstößen auf unterschiedlichem Abstraktionsniveau zu leisten. Die Nutzer profitieren voneinander durch Beiträge anderer Nutzer, die sie selbst nicht hätten geben können.

Beispiel 1: Sieht ein Nutzer nur, dass Daten weitergegeben werden, so identifiziert ein anderer Nutzer gegebenenfalls Empfänger im nicht-EU Ausland mit einem Datenschutzniveau, dass nicht dem der EU-Mitgliedsstaaten entspricht. Noch ein anderer Nutzer weiß, dass der Empfänger der Daten in den USA sitzt und personenbezogene Daten gemäß des Safe-Harbor Agreements, einer Übereinkunft, dass das Unternehmen bestimmte EU-Standards einhält, behandelt.

Gleichzeitig erlaubt es der Ansatz auch Unternehmen, sich selbst nach einheitlichen

[8] Interview TAZ, 03. August 2009

Bewertungskriterien strukturiert und systematisch auf mögliche Verstöße hin zu untersuchen. Den Behörden wiederum entsteht ein Vorteil dadurch, dass sie ihre beschränkten Ressourcen auf solche Anbieter konzentrieren können, für die interessierte Nutzer bereits Verstöße identifiziert haben.

Das entwickelte PET ist im Web frei zugänglich. Wir haben in einer Nutzerstudie mit dem PET und realen Anbietern gezeigt, dass Nutzer ohne Datenschutzkenntnisse 81 Prozent der Verstöße identifizieren können, die auch Experten finden – ein unserer Ansicht nach vielversprechender Ansatz, um dem Vollzugsdefizit und damit dem beschriebenen Nutzer-Anbieter-Problem entgegenzuwirken.

1.6.4. Anonymisierung als Ausweg für Nutzer und Anbieter

Im letzten Teil der Arbeit führen wir die Erkenntnisse aus den Studien zum Schutzbedürfnis der Nutzer untereinander als auch gegenüber dem Anbieter zusammen. Insbesondere berücksichtigen wir dabei auch die Interessen unterschiedlicher Parteien an den personenbezogenen Daten.

Wir greifen den Wunsch der Nutzer nach Anonymisierung bei kollaborativen Suchmaschinen auf. Anonymisierung bedeutet eine Abwägung zwischen Privatheit und dem verbleibenden Nutzen der Daten nach der Anonymisierung. Während sich existierende Arbeiten auf die Privatheit konzentriert haben, haben wir die Nutzenseite der Abwägung analysiert. Wir haben *erstens* aus Sicht der CSE-Nutzer untersucht, wie stark Anonymisierung die Menge der zwischen Nutzern austauschbaren Suchterme reduziert und für wie viele Nutzer eine Anonymisierung ihrer Suchanfragen überhaupt möglich ist. Anonymisierung ist *zweitens* auch aus Sicht des Anbieters von Interesse, da anonyme Daten nicht der Datenschutzgesetzgebung unterworfen sind. Somit würden Anbieter viele der oben identifizierten Verstöße nicht begehen. Damit ein Anbieter bereit ist, personenbezogene Daten zu anonymisieren, muss sichergestellt sein, dass er weiterhin seine ökonomischen Interessen verfolgen kann. Da Werbung die Haupteinnahmequelle von Suchmaschinenbetreibern ist und Nutzer nach unseren Erkenntnissen insbesondere Bedenken bezüglich personalisierter Werbung haben, untersuchen wir, ob ein Suchmaschinenanbieter anonymisierte Suchhistorien zum Schalten von Werbung einsetzten kann.

Wir können zeigen, dass Nutzer einer kollaborativen Suchmaschine, abhängig vom Grad der Anonymisierung, 18 bis 45 Prozent aller Suchterme auch anonymisiert austauschen können. Außerdem ist eine Anonymisierung von Suchhistorien für 70 bis 95 Prozent aller CSE-Nutzer möglich. Wir bewerten diese Zahlen positiv, berücksichtigt man, dass CSEs eigentlich gerade für den Austausch konkreter, also nicht anonymisierter Informationen konzipiert sind. Mit Blick auf den Anbieter verbleiben nach der Anonymisierung die Terme im Datensatz, die zu 61 bis 85 Prozent der möglichen Klicks auf Werbeanzeigen führen.

Insgesamt halten wir diese Ergebnisse für vielversprechend, um zum einen der Pri-

vatheit der Nutzer als auch zum andern den ökonomischen Interessen von Diensteanbietern gerecht zu werden.

Unsere Erkenntnisse aus den Studien zum Schutz der Privatheit zwischen Nutzern erlauben es den Entwicklern von Web 2.0-Diensten, Techniken zum Schutz der Privatheit deutlich besser auf die Privatheits- und Nutzungspräferenzen der Anwender hin zu entwickeln. Mit Hilfe des in dieser Arbeit geschaffenen Ansatzes zur Überprüfung von Datenschutzpraktiken von Anbietern können Unternehmen sich selbst, und die Nutzer sowie Datenschutzaufsichtsbehörden die Unternehmen, systematisch und zielgerichtet überprüfen. Zusammen mit der im letzten Teil vorgestellten Arbeit zum Nutzen anonymisierter Suchhistorien zeigen wir für alle – Nutzer, Diensteanbieter und Datenschutzaufsichtsbehörden – einen möglichen Ausweg aus dem Vollzugsdefizit auf.

1.7. Gliederung der Arbeit

In dem sich anschließenden Kapitel (Kapitel 2) vermitteln wir das erforderliche datenschutzrechtliche Wissen, auf das wir uns in den folgenden Kapiteln beziehen. Außerdem liefern wir Hintergrundinformationen zu dieser Arbeit und ordnen sie in den Kontext existierender Arbeiten ein. In Kapitel 3 beschreiben wir geleistete Vorarbeit. Dabei betrachten wir den Prozess einer Dienstnutzung von der Registrierung bis hin zur Löschung eines Zugangs sowie Besonderheiten bei der Nutzung von Web 2.0-Diensten. Kapitel 4 konzentriert sich darauf aufbauend auf den Aspekt des Nutzerverhaltens und deren Privatheitspräferenzen gegenüber anderen Nutzern. Kapitel 5.1 untersucht die Hypothese vom Vollzugsdefizit im Datenschutz. In Kapitel 5.2 stellen wir unser PET zur kollaborativen Identifikation von Datenschutzverstößen vor und evaluieren den Ansatz anhand einer Nutzerstudie. Kapitel 6 evaluiert die Möglichkeit der Nutzung anonymisierter Daten anhand von Suchhistorien. Kapitel 7 fasst die Ergebnisse zusammen und beschließt diese Arbeit mit einem kurzen Ausblick.

2. Grundlagen

Dieses Kapitel gibt einen Überblick über die rechtlichen als auch die technischen Grundlagen dieser Arbeit. Wir stellen in Abschnitt 2.1 Bedrohungen für die Privatheit vor. Diese und ähnliche Bedrohungen können entstehen, wenn, wie wir untersuchen werden, Daten unbewusst preisgegeben werden oder PETs nicht wirken. In Abschnitt 2.2 beschreiben wir die Entwicklung des Datenschutzes und das heute geltende, für diese Arbeit relevante Datenschutzrecht. In Abschnitt 2.3 betrachten wir Studien im Kontext Privatheit, Abschnitt 2.4 stellt existierende technische Ansätze zum Schutz der Privatheit vor.

2.1. Bedrohungen der Privatheit

Auch heute hört man immer noch die Aussage ‘Ich habe nichts zu verbergen – Datenschutz ist mir egal’ oder von Ansichten ähnlich der von Sun Microsystems’ CEO Scott McNealys ‘Du hast keine Privatheit – finde Dich damit ab’. Andere glauben, nichts preiszugeben und aus diesem Grund ‘privat’ zu sein. Ziel dieses Abschnitts ist es, mit solchen und ähnlichen Auffassungen aufzuräumen und exemplarisch aufzuzeigen, dass Datenschutzfragen jeden etwas angehen. Dazu führen wir zuerst den Begriff der digitalen Identität und mit ihr einhergehende Probleme ein. Anschließend beschreiben wir Privatheitsprobleme in Hinblick auf unsere Forschungsziele. Wir stellen am Ende jedes Absatzes den Bezug zu unserer Forschung her.

2.1.1. Digitale Identität und Privatheitsprobleme

[Una] versteht unter einer digitalen Identität ‘jede mögliche Form von technisch abgebildeten Daten, die zu einer Person gehören’. Das umfasst Daten zur Identifizierung, wie Nutzerkennungen und Pseudonyme, zur Authentifizierung, wie biometrische Daten, aber auch Merkmale wie Hobbys, Gewohnheiten etc. Die digitale Identität repräsentiert eine Person im Internet. Das birgt jedoch unter anderem folgende Gefahren:

Falschinformation Falsche Daten können zu einem verzerrten Bild einer Person führen, aber auch zu wirtschaftlichen Schäden, zum Beispiel bei der Vorhersage einer Kreditwürdigkeit. Falsche Daten zu korrigieren ist schwierig, da diese oftmals verteilt gespeichert werden, in Backups vorliegen etc.

Datenmissbrauch Daten können ohne Wissen und Einwilligung des Nutzers verkauft beziehungsweise an Dritte übermittelt werden. Der Nutzer kann solch einen Miss-

brauch erst erkennen, wenn er konkret von ihm betroffen ist, beispielsweise bei der Ablehnung eines Kredits, Jobs oder Ähnlichem.

Zweckentfremdung Daten werden zu einem bestimmten Zweck erhoben, den ein Anbieter seinen Nutzern mitteilen muss. Die Daten können jedoch zweckentfremdet werden, zum Beispiel zu Data-Mining-Zwecken oder für Werbung. Auch hier merkt der Nutzer dies nicht oder nur, wenn die Zweckentfremdung bereits stattgefunden hat.

Langfristige Aufbewahrung Daten können heute sehr lange aufbewahrt werden. Außerdem pflegen Nutzer das gleiche Profil in sozialen Netzwerkseiten über einen langen Zeitraum [Ros07]. Dadurch lassen sich Aussagen und Handlungen von Personen über lange Zeit nachvollziehen. Das ist problematisch, wenn die Selbstdarstellung der Vergangenheit, zum Beispiel als Schüler, deutlich von der heutigen Darstellung abweicht, beispielsweise als Arbeitssuchender.

Verknüpfbarkeit Das Verknüpfen von Daten unterschiedlicher öffentlicher und nichtöffentlicher (zum Beispiel unternehmensinterner) Informationsquellen kann dazu führen, dass individuell betrachtet harmlose Information in Kombination zu einem umfassenden Profil einer Person führt. Besonders problematisch kann das sein, wenn Daten falsch verknüpft werden, beispielsweise eine schlechte Zahlungsmoral mit dem Namen eines eigentlich pünktlich zahlenden Kunden.

Öffentliche Preisgabe Auch Daten, die Nutzer selbst über sich preisgeben, sind schützenswert. Das können soziale Kontakte sein, einmal getätigte Meinungen etc. Insbesondere kann auch eine unvollständige Preisgabe problematisch sein, zum Beispiel Informationen über den Kauf eines Medikamentes ohne eine exakte Angabe, ob die Medikamente überhaupt für den Käufer bestimmt sind.

2.1.2. Privatheitsbedrohungen Nutzer-Nutzer und Nutzer-Anbieter

Im Folgenden beschreiben wir exemplarisch Privatheitsbedrohungen, die entstehen, wenn (i) Nutzer Daten untereinander einsehen können, und Bedrohungen, (ii) die Anbieter durch die die Erhebung, Verarbeitung und Nutzung personenbezogener Daten schaffen.

Privatheitbedrohungen zwischen Nutzern Mögliche Konsequenzen aus einem unzureichenden Schutz der Privatheit zwischen Nutzern sind beispielsweise:

Identitätsdiebstahl Der Begriff Identitätsdiebstahl bezeichnet die unlegitimierte Nutzung einer fremden Identität. Dabei unterscheidet [Hoo07] zwei Formen des Missbrauchs: Bei der ersten legt der ‘Dieb’ eine neue Identität an, zu deren Validierung geklaute Daten eingesetzt werden. Bei der zweiten übernimmt der ‘Dieb’ einen Zugang bei einem Dienst. Dies kann beispielsweise durch Phishing erfolgen, bei dem Nutzer in gefälschten E-Mails aufgefordert werden, ihre Zugangsdaten preiszugeben. Identitäten können aber auch aus den Daten sozialer

Netzwerkseiten konstruiert werden. So kann ein 'Dieb' Name, Anschrift, Alter, etc. auslesen und sich damit bei einem Dienst anmelden.

Cyber- oder Online-Mobbing. Nach der Studie [LHU09] haben nahezu alle befragten Jugendlichen und jungen Erwachsenen in irgendeiner Form Erfahrungen mit Online-Mobbing gemacht, beispielsweise aufgrund 'peinlicher' Bilder. Nutzer stellen Daten über andere Nutzer online, die beschämend oder diskreditierend wirken.

Personalentscheidungen Laut einer Studie von Microsoft [Ser10] suchen 59 Prozent der Personalentscheider in Deutschland Informationen zu einem Bewerber im Internet. Wir betrachten den Personalentscheider hier ebenfalls als Nutzer, der Zugang zu beispielsweise einer sozialen Netzwerkseite hat. In den USA holen 79 Prozent der Personalentscheider regelmäßig oder sogar immer Informationen online ein. 16 Prozent der deutschen Personalentscheider haben Bewerber schon wegen unpassender Kommentare, Fotos oder Videos abgelehnt – und das mehrheitlich, ohne die gefundenen Informationen vorab nochmal auf Korrektheit zu prüfen.

Wir untersuchen in Kapitel 4, welche Privatheitspräferenzen Nutzer anderen Nutzern gegenüber haben. Außerdem analysieren wir den Einsatz von PETs zum Schutz zwischen Nutzern. Unsere Erkenntnisse sollen helfen, beschriebenen und ähnlichen Problemen entgegenzuwirken.

Privatheitsbedrohungen Nutzer–Anbieter Preisgegebene personenbezogene Daten können auch von Anbietern missbraucht werden. Personenbezogene Daten haben für viele Unternehmen einen Wert, das heißt einen Nutzen. Wir beschreiben im Folgenden anhand einer Auswahl prominenter Beispiele verschiedene Formen des Nutzens, das Privatheitsproblem und den Ursprung der Daten.

Personalisierte Werbung Der Umsatz, den Unternehmen wie Google (Google AdSense) durch Online-Werbung machen, hängt stark davon ab, wie gut sie die Interessen eines Nutzers identifizieren können [LSY03]. Bei Suchmaschinen wird Werbung durch die Auswertung von Anfragen und angeklickter Links personalisiert. Standortbezogene Dienste erlauben eine Personalisierung über Erkenntnisse, wo sich eine Person häufig aufhält oder gerade ist. So kann für diese gezielt solche Werbung geschaltet werden, die sich auf Geschäfte im nahen Umfeld bezieht. Das Privatheitsproblem für den Nutzer liegt dabei in dem Bilden der Profile. Diese sind für die eigentliche Diensterbringung oftmals nicht erforderlich und erlauben dem Anbieter weitgreifende Erkenntnisse über Interessen, Gewohnheiten, den Lebensmittelpunkt etc. des Nutzers. Die personenbezogenen Daten der Profile werden entweder von den Anbietern selbst erhoben oder zugekauft.

Scoring Unternehmen zur Kreditauskunft, wie die Schufa, nutzen personenbezogene Daten zur Erstellung eines sogenannten Scorings, das heißt, sie bewerten die Kre-

ditwürdigkeit einer Person. Diese Scorings können von anderen Unternehmen zum Beispiel im Falle einer anstehenden Finanzierung gegen eine Gebühr abgerufen werden. Offensichtlich helfen mehr Kenntnisse über eine Person bei der Bestimmung eines möglichst exakten Scores. Die Privatheit einer Person kann beispielsweise dann beeinträchtigt werden, wenn eine kreditwürdige Person als nicht kreditwürdig eingestuft wird, weil sie in einem sozialen Brennpunkt wohnt. Die Daten, die die Auskunfteien heranziehen, werden oftmals von denen an der Auskunft interessierten Unternehmen geliefert und anschließend von der Auskunftei zusammengeführt.

Preisdiskriminierung Eine weitere Konsequenz aus der Profilbildung ist Preisdiskriminierung. Das heißt, dass der gleiche Anbieter unterschiedlichen Kunden unterschiedliche Preise anbietet [Odl03]. Kenntnisse über eine Person erlauben es dem Anbieter, dessen Kaufkraft abzuschätzen oder auch sein Interesse an einem Produkt. Amazon.com stieß mit einem Versuch der Preisdiskriminierung auf großen Protest [BM02]. Offensichtlich möchte jeder Kunde den günstigsten Preis erhalten, unabhängig von seinem gesellschaftlichen Stand. Preisdiskriminierung und personalisierte Werbung werden oftmals gemeinsam betrachtet, das heißt, dass Werbung für unterschiedliche Kunden unterschiedliche Preise ausweist. Die Daten werden von den Unternehmen selbst erhoben oder zugekauft.

Versicherungstarife Versicherungen bieten unterschiedlichen Personen unterschiedliche Tarife und Preise an. Auch hier wäre es für den Anbieter von Interesse, viel über eine Person zu wissen, beispielsweise Essgewohnheiten bei einer Krankenversicherung. Um hier kein völlig zufälliges Preisgefüge entstehen zu lassen, die Privatheit als auch Personen vor Diskriminierung zu schützen, hat der Gesetzgeber Regeln erlassen, welche Informationen bei der Tarifgestaltung einfließen dürfen und welche nicht [Odl03]. Die Daten, anhand derer die Einordnung eines Kunden erfolgt, setzen sich aus Erfahrungen der Unternehmen und zugekaufte Daten zusammen.

Wir haben Szenarien beschrieben, in denen die Privatheit von Nutzern durch Unternehmen bedroht oder beeinträchtigt wird. Um die Privatheit der Nutzer zu schützen, definiert der Gesetzgeber Anforderungen an die Anbieter, die sie erfüllen müssen. Wir untersuchen in Kapitel 5.1, ob Anbieter im Internet diese Anforderungen erfüllen. Offensichtlich ziehen Anbieter einen Nutzen aus der Verwendung personenbezogener Daten und Nutzer profitieren davon, indem zum Beispiel personalisierte Werbung die Nutzung von Suchmaschinen und viele damit verbundenen Dienste (scheinbar) kostenlos macht. In Kapitel 6 untersuchen wir, ob es eine Abwägung zwischen Privatheit und Nutzen gibt, die die Anbieter wie die Nutzer zufriedenstellt.

Im weiteren Verlauf dieses Abschnitts beschreiben wir anhand von Beispielen aus der Presse und der Forschungsliteratur, wie Nutzer als auch Anbieter Daten bewusst oder unbewusst preisgeben, wie sie Daten erheben, speichern und kombinieren können

und warum die Löschung einmal preisgegebener Inhalte schwierig ist.

Unbewusste Datenpreisgabe Soziale-Netzwerkseiten wie Facebook und MySpace oder Xing und die VZ Netzwerke sind ausgesprochen beliebt. Millionen von Nutzern, 400 Millionen Nutzer alleine bei Facebook, von denen sich 50 Prozent jeden Tag anmelden [Fac10], geben auf diesen Seiten personenbezogene Informationen preis. Jeder Nutzer verfügt dabei über seine persönliche Profilseite, auf der er Inhalte bereitstellen sowie Verknüpfungen zu Freunden und Nutzergruppen erstellen kann. Zu den erzeugten Inhalten kann ein Nutzer seine Privatheitspräferenz definieren. Eine Beispielpräferenz ist 'zeige diese Bilder nur Freunden'. Viele Nutzer machen jedoch alle Inhalte öffentlich zugänglich, wenn auch Studien zeigen, dass dies oftmals aus Unwissenheit vor den Konsequenzen geschieht [Ros07, LHU09].

Wir untersuchen in Kapitel 3 und Kapitel 4 sowohl Privatheitspräferenzen als auch die unbewusste Preisgabe von Informationen genauer. Wir betrachten dazu Anbieter im Internet und insbesondere auch von unterschiedlichen Web 2.0-Technologien.

Privatheitsbewusste Nutzer Selbst die Personen, die auf ihre Privatheit achten, können schnell an das Licht der Öffentlichkeit gezerrt werden. Die Studie [ZG09] zeigt, dass Eigenschaften von Nutzern aufgrund von Gruppenzugehörigkeiten und Kontakten zu Freunden approximiert werden können. Und das, obwohl die Betroffenen die Privatheitspräferenz für ihr Profil auf 'privat' gestellt haben. Solch eine Vorhersage ist für das Attribut 'Geschlecht' mit einer Vorhersagegenauigkeit von über 73 Prozent möglich. Ein anderes Beispiel ist die Änderung der Privatheitseinstellungen, die Facebook an den Profilen seiner Nutzern durchgeführt hat: So wurden eigentlich als 'privat' eingestellte Profildaten plötzlich öffentlich zugänglich [sti10]. Außerdem hat Facebook versucht, ursprünglich private Nutzungsdaten wie Interessen an Inhalten an Dritte zu verkaufen [Aig10]. Solch eine Änderung des Zwecks der Nutzung der gespeicherten Daten ist ohne Einwilligung des Nutzers nicht zulässig. Durch den Aufkauf von Deja News, die von 1995 bis 2001 Newsgroupbeiträge gesammelt haben, hat Google einst vergessen geglaubte Informationen wieder für jeden zugreifbar gemacht [Hau01]. Heute sind diese historischen Beiträge unter dem Namen Google Groups[1] zu finden.

Geht ein Nutzer bewusst mit seinen Daten um, so kann er sich aufgrund der Praktiken der Anbieter trotzdem nicht sicher sein, dass seine Daten geschützt sind. Der Gesetzgeber schränkt den Zweck, zu dem personenbezogene Daten erhoben und genutzt werden dürfen, ein. Ohne Informierung und Einwilligung kann der Zweck nicht wie oben beschrieben willkürlich geändert werden. In Kapitel 5.1 untersuchen wir, inwieweit Anbieter gegen solche und ähnliche datenschutzrechtliche Anforderungen verstoßen.

[1] `http://groups.google.com`, Juni 2010

Indirekte Datenerhebung Auch Personen, die gar nicht über Onlineprofile oder Vergleichbares verfügen, werden von Internetdiensten erfasst. So kann der Nutzer einer sozialen Netzwerkseite dem Seitenbetreiber Zugriff auf seine E-Mailkonten gewähren und ihm erlauben, dort enthaltenen E-Mailadressen zum Verschicken von Einladungen zu der sozialen Netzwerkseite [sti10] zu nutzen. Dafür gibt der Nutzer – ob nun bewusst oder unbewusst – sein soziales Netzwerk in Form der Kontakte seiner E-Mailkorrespondenz an den Betreiber der sozialen Netzwerkseite preis [BGD+06]. Was er dabei sicherlich nicht berücksichtigt ist, dass er dabei auch die Privatheit seiner Kontakte gefährdet. Ein anderes Beispiel sind die Google E-Mailkonten: Selbst wenn der Besitzer eines solchen Kontos mit Datenschutzpraktiken wie 'Der Google Mail-Dienst präsentiert relevante Werbung [...] auf Grundlage [...] des Inhalts der Nachrichten'[2] einverstanden ist, ist mehr als fraglich, ob dies für Absender einer Nachricht an ein Google-Mailkonto ebenfalls gilt. Die Identifikation des Absenders ist aber durchaus möglich, zum Beispiel durch die Absender-IP-Adresse der E-Mail.

Diese indirekte Datenerhebung ist möglich, weil sich die Nutzer über den Informationsfluss innerhalb des Informationssystems nicht bewusst sind. Eines unserer Ziele ist die Identifikation von Privatheitspräferenzen von Nutzern. Damit diese sinnvoll erhoben werden können, müssen die Nutzer die Informationsflüsse verstanden haben. Wir schulen aus diesem Grund die Teilnehmer unserer Studien intensiv auf der eingesetzten Technologie (Kapitel 4).

Verknüpfung personenbezogener Daten Einmal preisgegebene Daten werden von Suchmaschinen indiziert und können leicht zu umfangreichen Profilen einer Person zusammengeführt werden. Besonders hervorzuheben sind dabei Personensuchmaschinen wie ´yasni.de' oder '123people.de'. Beispielsweise verknüpft 'yasni.de' zu einem Namen Informationen aus Online-Lexika, Branchenbüchern, Büchern, Dokumenten, Publikationen, Profilen sozialer Netzwerkseiten, Nachrichten und Forenbeiträge. Eine Verknüpfung kann aber auch von Daten innerhalb eines Unternehmens vorgenommen werden, zum Beispiel, wenn sich ein Nutzer mit unterschiedlichen Daten bei unterschiedlichen Diensten des gleichen Anbieters registriert hat. Nutzer können Informationen ebenfalls verknüpfen. Gibt ein Nutzer Daten einmal mit seinem realen Namen und Pseudonym sowie einmal nur mit dem Pseudonym preis, so lässt sich der Name der Person den pseudonymisierten Daten leicht zuordnen.

Wir werden in Kapitel 3 und Kapitel 4.2 analysieren, ob Nutzer die Gefahren der Verknüpfbarkeit von Informationen beim Einsatz von PETs berücksichtigen.

Speicherung, Weitergabe und Löschung Es gibt in der Bevölkerung die Auffassung [LHU09], dass man privatheitsgefährdende Inhalte 'ja jederzeit wieder löschen kann'. Dies ist aus mehrerer Hinsicht falsch: *Erstens*, mit einem Median von mehr als

[2] `http://mail.google.com/mail/help/intl/de/privacy.html`, Juli 2010

100 Freunden bei der sozialen Netzwerkseite MySpace [HJS] und tausender Freundesfreunde, können sehr schnell sehr viele Personen auf eine Information zugreifen. Unterstützt wird dies durch Techniken, die die Nutzer über neue Informationen auf den Seiten der Freunde informieren (engl. feeds). *Zweitens* wird eine Information, die einmal online war, vielerorts gespeichert, zum Beispiel in den Speichern von Suchmaschinen oder von Diensten wie archive.org. *Drittens* sind die betroffenen Personen nicht zwangsläufig auch die Personen, die die schützenswerte Information preisgegeben haben [SSP09]. Das Paradebeispiel sind Fotos, die im Allgemeinen nicht von der abgebildeten Person gemacht worden sind. Es ist also nicht möglich, eine Information einfach wieder zu löschen, sondern es sind Beschwerden beim Plattformbetreiber gegen die kritischen Inhalte erforderlich. *Viertens* ist zu berücksichtigen, dass wenn der Anbieter (oder andere Nutzer) die Daten bereits weitergegeben hat, der Betroffene fast keine Möglichkeiten mehr hat, den Fluss der Daten zu verfolgen. *Fünftens* werden viele Daten automatisiert im Hintergrund erhoben. Bis dem Betroffenen bewusst wird, dass unerwünschte Information über ihn kursiert, sind die Daten unter Umständen schon zwischen vielen Parteien ausgetauscht worden.

Wir untersuchen in Kapitel 5.1, ob zumindest zwischen Nutzer und Anbieter die dort geltenden gesetzlichen Vorschriften zum Auskunfts- und Löschersuchen eingehalten werden.

2.2. Datenschutzrecht

Dieser Abschnitt beleuchtet zuerst die historische Entwicklung des Datenschutzes auf internationaler und nationaler Ebene (Abschnitt 2.2.1). Anschließend beschreiben wir das für diese Arbeit, insbesondere Kapitel 5, relevante und heute geltende Recht, das heißt das Bundesdatenschutzgesetz (BDSG) (Abschnitt 2.2.2) und das Telemediengesetz (TMG) (Abschnitt 2.2.3).

Die Darstellung fokussiert auf die Aspekte, auf die die Folgekapitel aufbauen, das heißt, sie erhebt keinen Anspruch auf Vollständigkeit. Einen umfassenden Überblick liefern [TEG05, KSS08, Roß03, 9].

2.2.1. Entwicklung des Datenschutzes

Bei der Betrachtung der Entwicklung des Datenschutzes gehen wir zuerst auf internationale Abkommen und Gesetze ein. Anhand dieser Betrachtung wird deutlich, dass in vielen Ländern ein ähnliches Schutzniveau der Privatheit besteht wie in Deutschland. Das heißt, wir können davon ausgehen, dass unsere Ergebnisse aus Kapitel 5.1 und insbesondere aus Kapitel 5.2 mit leichten Anpassungen auch auf andere Länder übertragen werden können. Anschließend fokussieren wir auf die nationale Gesetzgebung in Deutschland seit dem ersten Datenschutzgesetz 1970. Wir beschreiben, was die

ursprüngliche Intention des Gesetzgebers gewesen ist und wie diese sich aus heutiger Sicht und dem Web 2.0 verändert hat.

OECD Eine der ersten internationalen Leitlinien mit entscheidendem Einfluss auf die heutige Datenschutzgesetzgebung kam 1980 von der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD). Sie trägt den Namen 'Leitlinien für den Schutz des Persönlichkeitsbereichs und den grenzüberschreitenden Verkehr personenbezogener Daten' [Org80]. Sie ist völkerrechtlich nicht bindend, das heißt, die Mitgliedsstaaten müssen diese nicht implementieren. Die in der Leitlinie definierten Grundsätze bei der Datenverarbeitung dienten aber bei vielen nationalen wie internationalen Gesetzen als Grundlage. Ihr Ziel war es, drei Grundsätze zu fördern: pluralistische Demokratie, Achtung der Menschenrechte und freie Marktwirtschaft [Org03]. Dabei sollte die Leitlinie die nationale Datenschutzgesetzgebung der OECD-Mitgliedsstaaten harmonisieren. Der grenzüberschreitende Fluss personenbezogener Daten war aus wirtschaftlicher Sicht erforderlich. Das Ergebnis der Harmonisierung sollte verhindern, dass nationale Gesetzgebungen diesen Fluss erschweren oder unterbinden, das heißt, dass sich Datenschutz nicht zu einem diskriminierenden Handelshemmnis entwickelt [KSS08, Seite 36]. Auf der anderen Seite sollte die Privatheit der betroffenen Personen nicht vernachlässigt werden. Die im Folgenden von der OECD aufgestellten Grundsätze lassen diese Verflechtung von Gesellschaftsordnung und wirtschaftlichen Fragen erkennen [Roß03, Kap. 2.3 Rn. 31]. Die Grundsätze sind:

Grundsatz der begrenzten Datenerhebung Dieser Grundsatz fordert eine Orientierung an der Erforderlichkeit der Daten zur Diensterbringung. Insbesondere soll die Sammlung nach geltendem Recht und 'fair' erfolgen. Das bedeutet beispielsweise, dass die Betroffenen von der Datenerhebung in Kenntnis gesetzt werden oder sogar explizit einwilligen müssen.

Grundsatz der Datenqualität Die Daten sollen sich auf den Zweck der Datenerhebung beziehen, aktuell, korrekt und vollständig sein.

Grundsatz der Zweckbestimmung Der Zweck der Datenerhebung muss vor der Datenerhebung kundgetan werden. Nach Erfüllung des Zwecks darf eine Nutzung nur noch zu Zwecken erfolgen, die mit dem ursprünglichen Zweck kompatibel sind.

Grundsatz der Nutzungsbegrenzung Daten sollen nur zu dem Zweck genutzt werden, zu dem sie erhoben wurden. Ausnahmen sind nur durch die Einwilligung des Nutzers möglich oder falls autorisiert durch ein Gesetz.

Grundsatz der Sicherung Die Daten sollen vor unautorisiertem Zugriff, Verlust, Zerstörung, Nutzung, Veränderung etc. geschützt werden.

Grundsatz der Offenheit Der Umgang mit personenbezogenen Daten soll offengelegt werden. Das heißt, dieser Grundsatz fordert den transparenten Umgang mit personenbezogenen Daten. Eine Beschreibung der Datenschutzpraktiken soll leicht

zugänglich sein, den Zweck der Datenerhebung / Verarbeitung / Nutzung wiedergeben sowie die Identität und den Sitz des Datenverarbeiters.

Grundsatz des Mitspracherechts Die Betroffenen sollen auch nach der Datenerhebung Zugriff auf ihre Daten haben. Sie sollen erfragen können, ob und welche Daten ein Unternehmen besitzt. Außerdem sollen die Betroffenen ihre Daten löschen, korrigieren, vervollständigen und ändern dürfen.

Grundsatz der Rechenschaftspflicht Es muss ein Verantwortlicher benannt werden, der bezüglich der Einhaltung der oben beschriebenen Grundsätze zur Rechenschaft gezogen werden kann.

Diese Richtlinie gilt für personenbezogene Daten sowohl im öffentlichen als auch im nicht-öffentlichen Bereich, durch die eine Gefährdung der Privatsphäre und Freiheit von Personen bewirkt werden kann [Org03].

Europarat Das Potential datenverarbeitender Technologien und die daraus resultierende Gefahr für die Privatheit der Bürger wurde vom Europarat früh erkannt. 1981 verabschiedete der Europarat das 'Übereinkommen zum Schutz des Menschen bei der automatischen Verarbeitung personenbezogener Daten' (Datenschutz-Konvention, Straßburger Vertrag) [Eur81]. Ziel war es, die 'grundlegenden Werte der Achtung des Persönlichkeitsbereichs und des freien Informationsaustausches zwischen den Völkern in Einklang zu bringen'. Das heißt, dass auch hier sowohl Vorgaben zu dem Umgang mit personenbezogenen Daten festgehalten wurden als auch Vorgaben zum freien Informationsfluss über Grenzen hinweg. Insbesondere wurde erkannt, dass personenbezogene Daten bei den Teilnehmerstaaten einem unterschiedlich hohen Schutzniveau unterliegen. Während die OECD-Leitlinie Datenschutz zu einem internationalen Thema gemacht hat, ist die herausragende Qualität dieses Übereinkommens die völkerrechtliche Verbindlichkeit für alle zeichnenden Staaten [KSS08]. Die Staaten müssen die Vorgaben ratifizieren und umsetzen.

Europäische Gemeinschaft (EG) EG-Recht ist supranationales Recht, das heißt, es muss in den nationalen Rechtssystemen der EG-Mitgliedsstaaten umgesetzt werden [TEG05]. Die Europäische Gemeinschaft hat seit 1975 Regelungen für die Verarbeitung personenbezogener Daten gefordert [KSS08]. 1995 erließ sie die EG-Datenschutzrichtlinie 95/46/EG (DSRL) [Par95]. Der Geltungsbereich der DSRL ist weit gefasst und umfasst 'alle personenbezogenen Daten, die in einer Datei gespeichert sind oder gespeichert werden sollen'. Das Herzstück der Datenschutzrichtlinie ist die Schaffung eines freien innereuropäischen Datenverkehrs. Insbesondere harmonisiert die Richtlinie die Datenschutzgesetzgebung zwischen den EG-Mitgliedsstaaten. Die Datenschutzrichtlinie kann auch als Umsetzung der OECD-Empfehlung betrachtet werden. Sie harmonisiert die Anforderungen an die Datenqualität (Art. 6 DSRL), die Zulässigkeit der Datenverarbeitung (Art. 7 DSRL), den Umgang mit sensiblen Daten (Art.

8 DSRL), Informationspflichten der verarbeitenden Stelle (Art. 10-11 DSRL) sowie ein Auskunfts- (Art. 12 DSRL) und Widerspruchsrecht (Art. 14 DSRL) des Betroffenen [KSS08]. Art. 16 und 17 machen Vorgaben zu der erforderlichen Vertraulichkeit beim Umgang mit personenbezogenen Daten. Artikel 22-24 DSRL konkretisieren die Ansprüche eines Betroffenen auf Rechtsbehelfe, eine Haftungsgrundlage, das heißt ein Anrecht auf die Forderung von Schadenersatz, und die Möglichkeit auf Sanktionen. Artikel 25 und 26 DSRL regeln die Übermittlung personenbezogener Daten in Drittländer.

Neben 95/46/EG seien zwei weitere Richtlinien erwähnt, die den Europäischen Datenschutz maßgeblich beeinflusst haben.

Erstens die Richtlinie 2002/58/EG 'über die Verarbeitung personenbezogener Daten und den Schutz der Privatsphäre in der elektronischen Kommunikation (Datenschutzrichtlinie für elektronische Kommunikation, EDSRL)'. Die EDSRL löste die Richtlinie 97/66/EG zur 'Verarbeitung personenbezogener Daten und den Schutz der Privatsphäre im Bereich der Telekommunikation' [Par97] ab. Ziel war es, eine Neuordnung des gemeinschaftlichen Telekommunikationsrechts zu schaffen. Insbesondere sollte die neue Richtlinie technikneutral sein – die alte Richtlinie orientierte sich stark an der Technik ISDN (Integrated Services Digital Network) – und offen für neue Entwicklungen [KSS08, Seite 64]. Die Richtlinie regelt das, was in Deutschland heute unter der geschäftsmäßigen Erbringung von Telekommunikationsdiensten gesehen wird.

Zweitens relevant ist 2006/24/EG, die 'Richtlinie über die Vorratsspeicherung von Daten, die bei der Bereitstellung öffentlich zugänglicher elektronischer Kommunikationsdienste oder öffentlicher Kommunikationsnetze erzeugt oder verarbeitet werden, und zur Änderung der Richtlinie 2002/58/EG' (VDSRL). Die Richtlinie entstand nach allgemeiner Sorge vor terroristischen Anschlägen, wie etwa nach dem Anschlag vom 11.09.2001. Die VDSRL schreibt die einheitliche und verdachtsunabhängige Speicherung von Kommunikationsdaten vor. Dies geschieht auf Vorrat, um auf die Daten im Einzelfall bei Ermittlung, Feststellung und Verfolgung schwerer Straftaten und Terrorismus zugreifen zu können [KSS08, Seite 67]. Anwendung findet die Richtlinie auf Verkehrs- und Standortdaten. Inhaltsdaten und Cookie-Daten sind von der Speicherung ausgenommen (siehe zu den unterschiedlichen Arten von Daten Abschnitt 2.2.2.2).

Mit Inkrafttreten des Vertrages von Lissabon im Dezember 2009 wurde die Europäische Union die Rechtsnachfolgerin der EG.

Vereinte Nationen Die Generalversammlung der Vereinten Nationen zum Schutz der Menschenrechte stellte 1990 Leitlinien für den Umgang mit der computergestützten Verarbeitung personenbezogener Daten auf [Uni90]. Die Ausgestaltung des Datenschutzes in dieser Leitlinie bleibt hinter den anderen beschriebenen Abkommen zurück [KSS08]. Es finden sich jedoch auch hier die Grundsätze zum Umgang mit personenbezogenen Daten und auch zur grenzüberschreitenden Weitergabe der Daten wie-

der.

Bundesrepublik Deutschland In Deutschland wurde 1970 das erste Datenschutzgesetz verabschiedet, genauer gesagt in Hessen (Hessisches Datenschutzgesetz, HDSG[3]). 1977 folgte das Bundesdatenschutzgesetz (BDSG) auf Bundesebene. Bemerkenswert an der Gesetzgebung ist, dass es bis zu diesem Zeitpunkt keine Datenskandale mit heutigem Ausmaß und heutiger Regelmäßigkeit gab. Die Initiative für das Gesetz beruhte rein auf Befürchtungen, wie aufkommende Technologien mit personenbezogenen Daten umgehen könnten [Roß03, Kap. 2.7 Rn. 2].

Ein Meilenstein in der Geschichte des Datenschutzes, mit der Einordnung des Datenschutzes als Grundrecht, entstand 1983 aus einer Protestbewegung gegen eine anstehende Volkszählung beziehungsweise gegen das Volkszählungsgesetz (VZG). Bei der Volkszählung sollten in Form einer Totalerhebung, bei der Beamte von Haus zu Haus gehen, personenbezogene Daten der Bürger erhoben werden. Insbesondere sollten für administrative Zwecke der Verwaltungen erhobene statistische Daten mit den Melderegistern abgeglichen werden (§9 Abs. 1 VZG). Über tausend Verfassungsbeschwerden [TEG05] stoppten das Vorhaben. Das Bundesverfassungsgericht (BVerGE) erkannte in einer die Beschwerden betreffenden Entscheidung, dem *Volkszählungsurteil* (BVerGE 65,1), Datenschutz als Grundrecht an. So heißt es in dem Urteil: 'Das Grundrecht gewährleistet insoweit die Befugnis des Einzelnen, grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten zu bestimmen'. Darüber hinaus erklärte es das gesamte Bundesgesetz für verfassungswidrig. Über den Stein des Anstoßes hinaus prüfte das Gericht außerdem die verfassungsrechtlichen Grundlagen des Datenschutzes beziehungsweise die des Rechts auf *informationelle Selbstbestimmung*. Das Recht auf informationelle Selbstbestimmung war ursprünglich als Abwehrrecht gegenüber hoheitlichem Handeln gedacht. Es leitet sich aus dem *allgemeinen Persönlichkeitsrecht* her, welches wiederum direkt im Grundgesetz (GG) verankert ist. Grundlage für das allgemeine Persönlichkeitsrecht ist das *Recht auf freie Entfaltung der Persönlichkeit* (Art. 2 Abs. 1 GG) und dem *unantastbaren Schutz der Menschenwürde* (Art. 1 Abs. 1 GG). Das Recht auf informationelle Selbstbestimmung bildet seit dem den Kern datenschutzrechtlicher Gesetzgebung. Unter den beschriebenen Einflüssen konnte erst 1987 die Volkszählung durchgeführt werden und das nur in modifizierter Form.

Seit 1970 wurde das Bundesdatenschutzgesetz dreimal novelliert [KB10]:

Die Arbeiten zur *BDSG-Novelle 1990* starteten bereits 1984 unter dem Eindruck des Volkszählungsurteils. Ziel war es nicht mehr nur Missbrauchsfälle zu verhindern. Vielmehr wollte der Gesetzgeber Grundsätze für einen 'fairen Umgang' mit personenbezogenen Daten etablieren [Roß03, Kap. 2.7 Rn. 40]. Außerdem wurde der steigenden Be-

[3] `http://www.datenschutz.hessen.de`, zuletzt gesichtet Juni 2010

deutung der privatwirtschaftlichen Verarbeitung personenbezogener Daten Rechnung getragen. So ist der Datenschutz nicht mehr nur Abwehr gegen den Staat, sondern vermehrt gegen privatwirtschaftliche Unternehmen. Entsprechend wurden in diesem Zuge auch die nicht-öffentlichen Stellen in die gesetzliche Regelung mit aufgenommen. Im September 1990 wurde die Gesetzesnovelle verabschiedet.

Die *BDSG-Novelle 2001* resultierte aus der Verpflichtung, die europäische Datenschutzrichtlinie 95/46/EG bis Oktober 1998 in nationales Recht umzusetzen. Die Umsetzung der DSRL brachte Grundsätze wie die Datenvermeidung und Datensparsamkeit, den Aspekt 'Datenschutz durch Technik' (Systemdatenschutz), pseudonymes und anonymes Handeln, Regeln zum Datenschutzaudit und die Selbstregulierung von Unternehmen mit ein. Neu waren auch Regelungen zur Videoüberwachung und zu mobilen Speicher- sowie Verarbeitungsmedien [Roß03, Kap. 2.7 Rn. 53]. Auch der grenzüberschreitende Datenfluss wurde vereinfacht. So gelten nach der Harmonisierung des Datenschutzrechtes zwischen den EU-Mitgliedsstaaten Anbieter innerhalb der EU nicht mehr als Dritte.

Die *BDSG-Novelle 2009* ist selbst wieder in drei Novellen unterteilt und umfasst über 90 Änderungen am BDSG (vgl. [4]). Die *Novelle I* ist seit April 2010 rechtskräftig. Ihr Schwerpunkt liegt auf den Rechten der Verbraucher, die von Auskunfteien und Scoring erfasst sind. Die *Novelle II* ist ein Gesetz zur Änderung datenschutzrechtlicher Vorschriften. Es stärkt die Position des Datenschutzbeauftragten, den Beschäftigten-Datenschutz, ändert die Regelungen zum Adresshandel und Werbung etc. Sie trat bereits im September 2009 in Kraft. Die *Novelle III* zum Verbraucherkredit wurde im Zuge einer Europäischen Verbraucherrichtlinie nötig und trat im Juni 2010 in Kraft. So müssen Auskunfteien nun auch europäische Anfragen beantworten und der Verbraucher muss über negative Auskünfte innerhalb der EU unterrichtet werden. Außerdem wurden neue Bußgeldvorschriften verabschiedet.

Während der Debatte um die zweite Novellierung des BDSG hat der Gesetzgeber neue Datenschutzgesetze für das Telekommunikations- und Medienrecht geschaffen. Diese zeichnen sich insbesondere durch ihre Kürze sowie präzise Vorgaben und Regelungen aus [Roß03, Kap. 2.7 Rn. 55]. Die für diese Arbeit relevanten Gesetze sind das Teledienstegesetz (TDG) und das Teledienstedatenschutzgesetz (TDDSG) des Bundes sowie der Mediendienste-Staatsvertrag (MDStV) der Länder. Diese Zweiteilung war ein politischer Kompromiss, nach dem die Länder die an die Allgemeinheit gerichteten Mediendienste regeln und der Bund die Dienste der Individualkommunikation (Teledienste) [KSS08, Seite 244]. Insbesondere zur Anpassung an die E-Commerce-Richtlinie 2000/31/EG [Par00] wurde 2001 das TDDSG aktualisiert und 2002 der MDStV entsprechend angepasst.

Die Zweiteilung (Mediendienste geregelt durch die Länder und Teledienste durch

[4] `http://www.securedataservice.de/Informationen/BDSGNovelle2009/tabid/221/Default.aspx` Juni 2010

den Bund) erwies sich jedoch als unvorteilhaft. Sie entsprach nicht der Ausgestaltung der Dienste im Realen, das heißt, durch die Vermischung von Medien- und Telediensten konnte auch die Verantwortlichkeit nicht klar zugeordnet werden. Vor diesem Hintergrund verständigte sich der Bund 2007 in der Verabschiedung des neunten Rundfunkänderungsstaatsvertrages und des Elektronischen-Geschäftsverkehr-Vereinheitlichungsgesetzes (ElGVG) auf die Aufhebung dieser Zweiteilung [KSS08, Seite 245]. Kern des ElGVG ist das Telemediengesetzes (TMG), das das bisherige TDG und das TDDSG ablöst. Interessant ist, dass das TMG keine Unterscheidung mehr vornimmt für öffentliche und private Anbieter von Telemedien. Für eine genaue Unterscheidung von Telemedien und Rundfunk, als auch von Telemedien und Telekommunikation verweisen wir auf [KSS08].

Die Auswirkungen des Volkszählungsurteils strahlen bis in die heutige Gesetzgebung ab. So können wir gespannt die nächstes Jahr (2011) auf EU-Ebene stattfindende Volkszählung erwarten. Unter dem Namen 'Zensus' basiert diese dabei nicht mehr auf einer Totalerhebung – 1983 stellte sich die Frage nach der Verhältnismäßigkeit einer Totalerhebung – sondern auf einer ausgefeilten Statistik zum Hochrechnen einer möglichst repräsentativen Auswahl (Sample) von Haushalten [Sta10].

2.2.2. Bundesdatenschutzgesetz (BDSG)

In diesem und dem folgenden Abschnitt stellen wir die heute geltenden und für diese Arbeit relevanten Merkmale des Bundesdatenschutzgesetzes und des Telemediengesetzes vor. Wir werden uns auf diese Merkmale jeweils beziehen, wenn wir in Kapitel 5 überprüfen, inwieweit sich Diensteanbieter im Internet konform zu geltendem Recht verhalten. Außerdem führen wir später verwendete Begrifflichkeiten, zum Beispiel zur Unterscheidung von unterschiedlichen Arten von Daten, von Anonymisierung und Pseudonymisierung etc., ein.

2.2.2.1. Aufbau des Bundesdatenschutzgesetzes

Das BDSG ist unterteilt in sechs Abschnitte. Der erste Abschnitt 'Allgemeine und gemeinsame Bestimmungen' umfasst die Begriffsbestimmung, die Grundregeln des Datenschutzes und die Anwendungsgebiete des Gesetzes. Der zweite Abschnitt befasst sich mit der 'Datenverarbeitung der öffentlichen Stellen', das heißt der Rechtsgrundlage, den Rechten der von der Datenerhebung Betroffenen und dem Bundesbeauftragten für den Datenschutz. Abschnitt drei, 'Datenverarbeitung nicht-öffentlicher Stellen und öffentlich-rechtlicher Wettbewerbsunternehmen', definiert erneut die Rechtsgrundlage und die Rechte des Betroffenen, hier jedoch für nicht-öffentliche Stellen. Außerdem werden die Rechte und Pflichten von Aufsichtsbehörden festgelegt. Abschnitt vier umfasst 'Sondervorschriften', zum Beispiel bei der Datenverarbeitung in der Forschung.

Abschnitt fünf beschreibt ‘Schlussvorschriften’ mit Bußgeld- sowie Strafvorschriften. Abschnitt sechs befasst sich mit ‘Übergangsvorschriften’ für Daten, die vor der Novelle 2001 erhoben wurden.

2.2.2.2. Unterscheidung von Daten

In Paragraph §3 definiert das BDSG die wichtigsten Begriffe. Es kennt vier unterschiedliche Formen von Daten. Die erste Form, und zugleich der Anknüpfungspunkt des BDSG, umfasst *personenbezogen Daten* einer natürlichen Person (§3 Abs. 1). Dies ‘sind Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person’. Eine Unterteilung der personenbezogenen Daten wird durch die Unterscheidung in *sensible* Daten (§3 Abs. 9) vorgenommen. Sensible Daten sind ‘Angaben über die rassische und ethnische Herkunft, politische Meinungen, religiöse oder philosophische Überzeugungen, Gewerkschaftszugehörigkeit, Gesundheit oder Sexualleben.’

Anfallen können diese Daten in Form von *Bestandsdaten*, *Verkehrsdaten*, *Nutzungsdaten* oder *Inhaltsdaten* [TEG05].

Bestandsdaten Bestandsdaten werden in der Regel vor einer Dienstnutzung erhoben. Sie identifizieren eine Person, ordnen beispielsweise einen Telefon- oder Internetanschluss einer Person zu.

Verkehrsdaten Verkehrsdaten entstehen bei der *Erbringung* eines Dienstes. Beim E-Mail-Austausch sind dies Daten wie Absender und Empfänger.

Nutzungsdaten Nutzungsdaten fallen bei der *Nutzung* eines Dienstes an, zum Beispiel bei der Identifikation zur Anmeldung auf einer Webseite. Nutzungsdaten dürfen nur zu Abrechnungszwecken gespeichert werden.

Inhaltsdaten Hierbei handelt es sich um bei der Dienstnutzung ausgetauschte Informationen, das heißt den Inhalt von E-Mails, von Telefonaten etc.

Von besonderer Relevanz für diese Arbeit ist die Unterscheidung zwischen den bereits beschriebenen *personenbezogenen* Daten und *pseudonymisierten* respektive *anonymisierten* Daten, auf die wir im Folgenden eingehen.

Zur Erklärung helfen die Begriffe *bestimmte* und *bestimmbare* Person. Eine bestimmte Person ist bestimmt, wenn sie mit den Angaben, zum Beispiel aus einem Datensatz, eindeutig identifizierbar ist. Dies kann der Name sein, ein namentlich beschriftetes Bild in der Medizin, etc. Bestimmbar ist eine Person, wenn durch die Verknüpfung von Hintergrundwissen, auch als korrelierendes Wissen bezeichnet, der Personenbezug hergestellt werden kann.

Pseudonymisierung ist im BDSG §3 Abs. 6a wie folgt definiert:

> *Pseudonymisieren* ist das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren.

Eine Instanz, die Daten pseudonymisiert, ersetzt also die Merkmale, die eine Person eindeutig bestimmen. Beliebt ist dieser Ansatz beispielsweise bei Einträgen in Internetforen. Hier wird zumeist nicht der echte Name des Autors angezeigt, sondern ein Alias. Das Problem dabei ist, dass nahezu kein Schutz davor entsteht, Hintergrundwissen zu verknüpfen. Das heißt, Attribute, die bei der Pseudonymisierung nicht ersetzt wurden, können unter Umständen zu der Verknüpfung herangezogen werden und eine Person bestimmen. [Swe00] und [Gol06] haben gezeigt, dass die Krankenakten von 63 bis 87 Prozent der amerikanischen Bevölkerung durch die Kombination der Attribute Postleitzahl, Geburtsdatum und Geschlecht eindeutig einer Person zugeordnet werden können. Dies ist möglich, da der zugehörige Name über die amerikanischen Wählerlisten herausgefunden werden kann. Diese umfassen neben dem Namen ebenfalls die Attribute Postleitzahl, Geburtsdatum und Geschlecht, über deren Verknüpfung mit den Krankendaten eine Person identifiziert werden kann. Die Pseudonymisierung des Namens wäre also an dieser Stelle ohne Wirkung für die Privatheit. Attribute oder deren Kombination, die einen Datensatz eindeutig identifiziert, nennt [Swe02] *Quasi-Identifkatoren*.

Das Prinzip der Anonymisierung versucht die Korrelation von Wissen zu erschweren. Das BDSG definiert Anonymisierung BDSG §3 Abs. 6 als:

> *Anonymisieren* ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können.

Eine 100% sichere Anonymisierung gibt es nicht, da dies erfordern würde, dass sowohl alles heute verfügbare Hintergrundwissen, als auch das von morgen berücksichtigt würde. Wie auch schon Formulierungen wie ‘nur mit einem unverhältnismäßig großen Aufwand’ erkennen lassen, ist Anonymisierung daher eine Frage der Wahrscheinlichkeit [TEG05], dass eine Person identifiziert werden kann. Wir werden in Abschnitt 2.4.2 unterschiedliche Anonymisierungsverfahren für unterschiedliche Szenarien und Technologien vorstellen.

Die Relevanz der Mechanismen der Pseudonymisierung und Anonymisierung ergibt sich direkt aus dem Ansatzpunkt des BDSG, das heißt den personenbezogenen Daten. Ist ein Personenbezug nicht oder nicht mehr herstellbar, so greifen auch die Regelungen des BDSG nicht [TEG05].

2.2.2.3. Erhebung, Verarbeitung und Nutzung von Daten

Das BDSG unterscheidet drei grundlegende Aktionen beim Umgang mit personenbezogenen Daten: die Erhebung, die Verarbeitung und die Nutzung.

Erhebung 'Erheben ist das Beschaffen von Daten über den Betroffenen' (BDSG §3 Abs. 3) Die Erhebung soll in erster Linie bei dem Betroffenen selbst stattfinden und fair erfolgen, das heißt, der Betroffene soll die Folgen der Datenpreisgabe abschätzen können.

Verarbeitung Die Verarbeitung personenbezogener Daten ist ein Sammelbegriff für das *Speichern*, *Verändern*, *Übermitteln*, *Sperren* und *Löschen* von Daten.

Speichern Speichern bezeichnet das 'Erfassen, Aufnehmen oder Aufbewahren personenbezogener Daten auf einem Datenträger zum Zwecke ihrer weiteren Verarbeitung oder Nutzung' (BDSG §3 Abs. 4 Nr. 1). Damit das BDSG Anwendung findet, muss das Ziel der Speicherung die weitere Verarbeitung oder Nutzung sein beziehungsweise eine Verwendungsabsicht vorliegen.

Verändern Verändern ist 'das inhaltliche Umgestalten gespeicherter personenbezogener Daten.' (BDSG §3 Abs. 4 Nr. 2) Ändern bezeichnet dabei nicht nur den Austausch von Attributwerten, zum Beispiel die Änderung des Attributs' Familienstand' von 'ledig' auf 'verheiratet'. Vielmehr umfasst Verändern jegliche Veränderungen, die einen neuen Aussagewert schaffen. Das schließt somit auch jede Form des Einfügens und Entfernen von Daten mit ein.

Übermitteln Eine besondere Bedrohung der Privatheit einer Person entsteht durch einen ungehinderten Austausch personenbezogener Daten, das heißt die Übermittlung an Dritte. Eine Übermittlung in Drittstaaten birgt vielfach zusätzlich die Gefahr, dass eine Rechtsgrundlage im Empfängerland gilt, die kein ebenbürtiges Schutzniveau der Privatheit des Betroffenen garantiert. Wir betrachten den Fall der Datenweitergabe explizit in Abschnitt 2.2.2.4.

Sperren Sperren bezeichnet 'das Kennzeichnen gespeicherter personenbezogener Daten, um ihre weitere Verarbeitung oder Nutzung einzuschränken'. Es stellt eine Alternative zum Löschen dar, wenn beispielsweise die personenbezogenen Daten nach einem Löschersuchen des Betroffenen aufgrund gesetzlicher Vorschriften weiter gespeichert werden müssen.

Löschen Löschen ist 'das Unkenntlichmachen gespeicherter personenbezogener Daten'. Löschen erfordert das Vernichten des Datenträgers, das Entfernen personenbezogener Angaben durch Schwärzen oder das mehrfache Überschreiben bei der elektronischen Datenverarbeitung.

Nutzen Nutzen von Daten ist 'jede Verwendung personenbezogener Daten' (BDSG §3 Abs. 5). Eine Nutzung liegt aber nur dann vor, wenn sich diese auch auf Personen bezieht. Im Unterschied zum Ändern von Daten entsteht bei deren Nutzung kein neuer Aussagewert.

2.2.2.4. Verantwortliche Stelle / Empfänger / Dritter

Wie im vorangegangenen Abschnitt beschrieben, stellt die Weitergabe, das heißt die Übermittlung personenbezogener Daten, eine besondere Herausforderung für die Gesetzgebung dar.

Die Grundlage für den Gesetzgeber besteht in der Unterscheidung der *verantwortlichen Stelle*, des *Empfängers* und *Dritter*.

verantwortliche Stelle Nach BDSG §3 Abs. 7 ist die verantwortliche Stelle (i) 'jede Person oder Stelle', (ii) 'die personenbezogene Daten für sich selbst erhebt, verarbeitet oder nutzt' oder (iii) 'dies durch andere im Auftrag vornehmen lässt.' Insbesondere gibt es kein Konzernprivileg, das heißt, im rechtlichen Sinne sind Unternehmen nach §15 Aktiengesetz (AktG) eines Konzernverbundes eigene verantwortliche Stellen [TEG05]. Die verantwortliche Stelle ist immer die Firma einschließlich ihrer Niederlassungen und Zweigstellen. Ist die Zweigstelle allerdings außerhalb der EU und des EWR ist auch eine Zweigstelle immer Dritter.

Empfänger Gemäß dem BDSG ist ein Empfänger 'jede Person oder Stelle, die Daten erhält.' Eine Weitergabe von Daten muss jedoch nicht zwangsläufig an einen Dritten erfolgen. Neben dem gerade beschriebenen Fluss von Daten innerhalb der verantwortlichen Stelle ist das auch dann der Fall, wenn der vermeintliche Dritte derjenige ist, auf den sich die Daten beziehen.

Dritte Ein 'Dritter ist jede Person oder Stelle außerhalb der verantwortlichen Stelle' (BDSG §3 Abs. 8). Weiter gilt die Ausnahme 'Dritte sind nicht der Betroffene sowie Personen und Stellen, die im Inland, in einem anderen Mitgliedsstaat der Europäischen Union oder in einem anderen Vertragsstaat des Abkommens über den Europäischen Wirtschaftsraum personenbezogene Daten im Auftrag erheben, verarbeiten oder nutzen.'

2.2.2.5. Verbot mit Erlaubnisvorbehalt

Das grundlegende Prinzip im BDSG ist das *Verbot mit Erlaubnisvorbehalt* (§4 Abs. 1 BDSG). Es besagt, dass die Erhebung, Verarbeitung und Nutzung generell verboten ist, außer dass (i) ein anderes Gesetz den Anbieter legitimiert oder auffordert, (ii) das BDSG es explizit erlaubt oder (iii) der Betroffene einwilligt.

Gesetzliche Legitimierungen Außerhalb des BDSG definierte Rechtsvorschriften können das Erheben, Verarbeiten und Nutzen personenbezogener Daten erlauben oder anordnen [KSS08, Seite 132]. Für diese Arbeit ist das insbesondere das TMG. Es gilt der Grundsatz, dass die *speziellere Norm* der allgemeineren Norm vorgeht, im Kontext der Telemedien also das Telemediengesetz (TMG). Das BDSG stellt dabei den Charakter des *Auffanggesetzes* dar, das heißt, ist in keiner speziellen Norm ein Sachverhalt

explizit geregelt, greift das BDSG.

Das BDSG regelt daher die Erhebung, Verarbeitung und Nutzung ebenfalls. Um nur ein Beispiel zu nennen, geschieht dies im Kontext der Sonderregelungen bei der Erhebung, Verarbeitung und Nutzung personenbezogener Daten zu Forschungszwecken.

Einwilligung Für die Einwilligung gelten besondere Anforderungen: Nach BDSG §4a Abs. 1 ist die 'Einwilligung [...] nur wirksam, wenn sie auf der freien Entscheidung des Betroffenen beruht.' Die Einwilligung ist eine Willenserklärung und muss *bewusst* erfolgen. Eine Einwilligung muss zum richtigen *Zeitpunkt*, das heißt vor der Erhebung, Verarbeitung oder Nutzung personenbezogener Daten erfolgen. Außerdem ist sicherzustellen, dass der Betroffene einsichtsfähig, das heißt zum Beispiel alt genug zum Verstehen der Einwilligungserklärung, ist. Der Betroffene ist außerdem über die *Art* der Daten, auf die sich die Einwilligung bezieht, den *Zweck* und den *Umfang* der Einwilligung zu unterrichten.

Eine Einwilligung kann auch elektronisch erklärt werden, wenn der Anbieter sicherstellt, dass (i) die Einwilligung eine eindeutige und bewusste Handlung des Nutzers ist, sie (ii) protokolliert wird und (iii) der Inhalt der Einwilligung jederzeit von dem Betroffenen oder dem Nutzer abgerufen werden kann (§28 Abs. 3a). Die Einwilligung darf mit anderen Erklärungen abgegeben werden. In einem solchen Fall ist die Einwilligung besonders hervorzuheben [9].

Eine Einwilligung muss 'jederzeit mit Wirkung für die Zukunft widerrufen' werden können (BDSG §28 Abs. 3a). Handelt es sich um eine elektronisch gegebenen Einwilligung, sollte dies ohne Medienbruch möglich sein [9]. So sollte eine elektronisch gegebene Einwilligung auch möglichst elektronisch widerrufbar sein.

2.2.2.6. Datenschutzkontrolle

Die Datenschutzkontrolle unterteilt sich in zwei Bereiche [KSS08]: Die Fremdkontrolle und die Selbstkontrolle.

Die Fremdkontrolle findet auf Bundesebene durch den 'Datenschutzbeauftragten für den Datenschutz und die Informationsfreiheit' statt (§§22 ff. BDSG). Auf Landesebene erfolgt sie für den öffentlichen Bereich entsprechend durch den 'Landesbeauftragten für den Datenschutz'. Außerdem gibt es Aufsichtsbehörden für den nicht-öffentlichen Bereich (§22 BDSG), also den Bereich, der für unsere Arbeit von besonderem Interesse ist. Der nicht-öffentliche Bereich fällt in den Zuständigkeitsbereich der Länder. Dabei ist die Verantwortung teils den Innenresorts, den Regierungsbezirken, dem Landesverwaltungsamt, der Aufsichts- und Diensleistungsdirektion oder dem Landesdatenschutzbeauftragten zugeordnet. Die Aufsichtsbehörden kontrollieren von Amts wegen, es bedarf also nicht der Initiative Dritter oder eines konkreten Anlasses. Die Unternehmen müssen mit den Aufsichtsbehörden kooperieren. Nach der Novelle II des

BDSG können die Aufsichtsbehörden direkt gegen Verstöße vorgehen, beispielsweise eine Praktik untersagen. Vorher konnten sie nur organisatorische oder technische Maßnahmen auferlegen [KB10]. Wie wir in Kapitel 5 zeigen werden, ist die effektive Kontrolle tausender Dienste mit den verfügbaren Mitteln nicht möglich. Wir werden ein PET entwickeln, dass die Behörden bei einer effektiven und effizienten Kontrolle unterstützt.

Die Selbstkontrolle umfasst Meldepflichten (§§4d, e BDSG), Vorabkontrollen (§4d Abs. 5 BDSG) und behördliche beziehungsweise betriebliche Beauftragte für den Datenschutz. Meldepflichten erfordern, dass Unternehmen (hier nur auf den nicht-öffentlichen Bereich bezogen) Verfahren zur automatisierten Verarbeitung personenbezogener Daten den Aufsichtsbehörden melden. Diese Pflicht entfällt, wenn das Unternehmen einen Verantwortlichen für den Datenschutz stellt. Vorabkontrollen erfolgen vor der Inbetriebnahme des datenverarbeitenden Dienstes. Das findet insbesondere bei der Verarbeitung besonders sensibler Daten statt oder wenn die Bewertung einer Persönlichkeit erfolgen soll. Der betriebliche Beauftragte für den Datenschutz ist Teil der verantwortlichen Stelle und soll die Einhaltung der rechtlichen Vorgaben kontrollieren, insbesondere beim Einsatz von Datenverarbeitungsprogrammen. Wir werden in Kapitel 5 zeigen, dass auch die betrieblichen Datenschutzbeauftragten überfordert sind und einen Mechanismus brauchen, um die Datenschutzpraktiken des eigenen Unternehmens zu überprüfen.

2.2.3. Telemediengesetz (TMG)

Das Telemediengesetz ist ein bereichsspezifisches Datenschutzrecht. Es wurde erforderlich durch die allgegenwärtige Verarbeitung personenbezogener Daten im Internet und insbesondere des Web 2.0 sowie wegen der Durchdringung des Alltages mit unterschiedlichen Technologien und Anwendungen. So finden Einkäufe heute vielfach online statt und das oftmals von mobilen Endgeräten aus. Außerdem wird die klassische Dienstnehmer / Dienstgeber-Architektur (Client / Server-Architektur) zunehmend von Rechner-zu-Rechner-Systemen (Peer-to-Peer) verdrängt. Weiter ist bei serviceorientierten Architekturen die Darstellung von den Inhalten getrennt. Mit der Nutzung der Telemediendienste hinterlassen die Anwender Datenspuren, die zu umfassenden Profilen über Gewohnheiten, Interessen, etc. führen können. Das TMG versucht Transparenz zu schaffen, das heißt dem Nutzer einen vertrauenswürdigen Umgang mit den personenbezogenen Daten zu gewährleisten.

Das Telemediengesetz besteht aus fünf Abschnitten: Erstens ‘Allgemeine Begriffsbestimmungen’, zweitens ‘Zulassungsfreiheiten und Informationspflichten’, drittens ‘Verantwortlichkeit’, viertens ‘Datenschutz’ und fünftens ‘Bußgeldvorschriften’. Wir werden nur die für diese Arbeit relevante Auszüge aus dem TMG vorstellen. Weiterführende Informationen geben [KSS08, 9].

2.2.3.1. Begriffsbestimmungen

Ein *Diensteanbieter* im Sinne des TMG ist ‘jede natürliche oder juristische Person, die eigene oder fremde Telemedien zur Nutzung bereithält oder den Zugang zur Nutzung vermittelt’ (§2 Abs. 1 TMG). Der *Nutzer* eines Telemediendienstes ist ‘jede natürliche oder juristische Person, die Telemedien nutzt, insbesondere um Informationen zu erlangen oder zugänglich zu machen’ (§2 Abs. 3). Das *Sitzland* eines Telemedienanbieters bestimmt sich danach, ‘wo dieser seine Geschäftstätigkeit tatsächlich ausübt. Dies ist der Ort, an dem sich der Mittelpunkt der Tätigkeiten des Diensteanbieters im Hinblick auf ein bestimmtes Telemedienangebot befindet’ (§2a Abs. 1). Das *Herkunftslandprinzip* legt in §3 Abs. 1 fest, dass die Anwendbarkeit des deutschen Rechts davon abhängt, wo ein Anbieter ‘seine Geschäftstätigkeit tatsächlich ausübt. Dies ist der Ort, an dem sich der Mittelpunkt der Tätigkeiten des Diensteanbieters im Hinblick auf ein bestimmtes Telemedienangebot befindet’.

Bestandsdaten Bestandsdaten sind Daten, die für die ‘Begründung, inhaltliche Ausgestaltung oder Änderung eines Vertragsverhältnisses zwischen dem Diensteanbieter und dem Nutzer über die Nutzung von Telemedien erforderlich sind’ (§14 Abs. 1 TMG). Bestandsdaten sind also die Grunddaten des Vertragsverhälnisses [KSS08, Seite 257] wie Name, Anschrift, Geburtsdatum, Bank- und Zahlungsdaten. Sie dürfen vom Anbieter ohne Einwilligung des Nutzers erhoben werden, soweit sie zur Diensterbringung unerlässlich sind. Erlässlich wäre es beispielsweise, mehrere Kontaktdaten (E-Mail und Telefon) zu erheben. In diesem Fall müsste der Nutzer in die Erhebung der erlässlichen Daten einwilligen.

Nutzungsdaten Nutzungsdaten sind für die Inanspruchnahme von Telemedien und deren Abrechnung erforderlich. Dies umfasst insbesondere Merkmale zur Identifikation des Nutzers, Angaben über Beginn, Ende und Umfang der Nutzung und Angaben über die vom Nutzer in Anspruch genommenen Telemedien.

Zweckbindungsgrundsatz Der Zweckbindungsgrundsatz (§12 Abs. 2 TMG) ist einer der wichtigsten Bestandteile des TMG. Er sagt aus, dass ‘für die Bereitstellung von Telemedien erhobene personenbezogene Daten für andere Zwecke nur [verwendet werden dürfen], soweit dieses Gesetz oder eine andere Rechtsvorschrift, die sich ausdrücklich auf Telemedien bezieht, es erlaubt oder der Nutzer eingewilligt hat.’ Dabei muss sich die Rechtsvorschrift auch auf die Art der jeweiligen Daten (Bestandsdaten, Nutzungsdaten) beziehen. So gelten wie beschrieben die Erlaubnistatbestände für Nutzungsdaten (Abschnitt 2.2.3.2) nicht gleichermaßen für die Bestandsdaten.

2.2.3.2. Erlaubnistatbestände zur weiteren Verwendung von Nutzungsdaten

Abrechnungszwecke Soweit Nutzungsdaten 'für Zwecke der Abrechnung mit dem Nutzer erforderlich sind' (§15 Abs. 4) darf der Diensteanbieter auch über die Nutzungsdauer hinaus speichern. Nutzungsdaten über unterschiedliche Telemedien dürfen zu Abrechnungszwecken außerdem zusammengeführt werden (§15 Abs. 2)

Pseudonymisierte Nutzungsprofile Nutzungsprofile sind Informationssammlungen über beispielsweise das Einkaufsverhalten einer Person in einem Online-Shop. Nutzungsprofile erlauben, abhängig vom Kontext, Rückschlüsse auf Gewohnheiten, Vorlieben, Interessen, etc. von Nutzern. Oftmals werden Nutzungsprofile auch über lange Zeit erstellt. Zum 'Zwecke der Werbung, der Marktforschung oder zur bedarfsgerechten Gestaltung der Telemedien' (§15 Abs. 3) darf ein Telemedienanbieter Nutzungsprofile erstellen, solange diese auf Pseudonymen beruhen. Der Nutzer hat das Recht, der Erstellung dieser Profile zu widersprechen, und er ist auf diese Möglichkeit hinzuweisen.

Anonymisierte Daten Pseudonymisierte Nutzungsprofile dürfen ohne Einwilligung des Nutzers nicht übermittelt, zu anderen Zwecken verwendet oder mit anderen Daten verknüpft werden [KSS08, Seite 259]. Sind die Nutzungsdaten hingegen anonymisiert, dürfen sie zum Zwecke der Marktforschung anderen Diensteanbietern übermittelt werden (§15 Abs. 5).

Missbrauchsbekämpfung Ein Diensteanbieter kann zum ´Zweck der Rechtsverfolgung' (§15 Abs. 8) von Missbrauchsfällen Nutzungsdaten auch über den ursprünglichen Zweck hinaus verwenden. Dies ist möglich, wenn es dokumentierte Anhaltspunkte gibt, dass ein Nutzer das Entgelt für einen Telemediendienst nicht oder nur teilweise zu entrichten plant.

Auch für Bestandsdaten gibt es Erlaubnistatbestände, die für diese Arbeit jedoch eine untergeordnete Rolle spielen. So darf gemäß §14 Abs. 2 ein 'Diensteanbieter im Einzelfall Auskunft über Bestandsdaten erteilen, soweit dies für Zwecke der Strafverfolgung, zur Gefahrenabwehr durch die Polizeibehörden der Länder, zur Erfüllung der gesetzlichen Aufgaben der Verfassungsschutzbehörden des Bundes und der Länder, des Bundesnachrichtendienstes oder des militärischen Abschirmdienstes oder des Bundeskriminalamtes im Rahmen seiner Aufgabe zur Abwehr von Gefahren des internationalen Terrorismus oder zur Durchsetzung der Rechte am geistigen Eigentum erforderlich ist.'

2.2.3.3. Pflichten des Anbieters

In diesem Abschnitt stellen wir aus dem TMG resultierende Pflichten für die Anbieter vor. Wir werden die Einhaltung dieser Pflichten insbesondere in Kapitel 5 untersuchen.

Informations- und Unterrichtungspflichten Nur ein informierter Nutzer, der nachvollziehen kann, wer, was, zu welchem Zeitpunkt über ihn weiß, ist in der Lage, sein Recht auf informationelle Selbstbestimmung wahrzunehmen [KSS08, Seite 264]. Zu diesem Zweck, das heißt zur Schaffung von Transparenz gemäß dem Grundsatz der Offenheit, ist der Anbieter verpflichtet, den Nutzer über den Umgang mit den personenbezogenen Daten zu informieren.

So muss der Anbieter den Nutzer nach §13 TMG '*zu Beginn* des Nutzungsvorgangs über *Art, Umfang und Zwecke* der Erhebung und Verwendung personenbezogener Daten' informieren. Weiter muss der Anbieter den Nutzer unterrichten, wenn 'die Verarbeitung seiner Daten in Staaten außerhalb des Anwendungsbereichs der Richtlinie 95/46/EG des Europäischen Parlaments und des Rates vom 24. Oktober 1995 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr' erfolgt. Insbesondere muss eine Unterrichtung in '*allgemein verständlicher Form*' erfolgen und 'jederzeit abrufbar sein'. Nach §13 Abs. 2 kann eine Einwilligung *elektronisch erklärt* werden, wenn der Nutzer sie bewusst und eindeutig erteilt. Für Telemedienanbieter hat sich die doppelte Einwilligung etabliert. Bei der doppelten Einwilligung setzt der Nutzer beispielsweise erst ein Häkchen (i) zur Einwilligung und bestätigt dies (ii) durch das Betätigen einer Schaltfläche (Knopf/Button) oder das Anklicken eines Links. Ist die Einwilligung Teil einer Kombinationserklärung [KSS08, Seite 264], also beispielsweise eine Kombination aus Unterrichtung und Einwilligung, so muss der Teil, in den eingewilligt wird, von der bloßen Unterrichtung unterscheidbar sein. Weiter muss die Einwilligung *protokolliert* werden und der Nutzer seine Einwilligung jederzeit mit Wirkung für die Zukunft *widerrufen* können. Der Anbieter muss auf das Widerrufsrecht *vor der Einwilligung* hinweisen. Bei der Weitervermittlung zu einem anderen Anbieter (§13 Abs. 5) muss der Anbieter dies dem Nutzer anzeigen. Kann ein Dienst *anonymisiert* oder unter Verwendung eines *Pseudonyms* genutzt werden, muss der Anbieter den Nutzer über diese Möglichkeit informieren (§13 Abs. 6 TMG).

Automatisierte Verfahren Setzt ein Anbieter automatisierte Verfahren ein, die 'eine spätere Identifikation des Nutzers ermöglich[en] und eine Erhebung oder Verwendung personenbezogener Daten vorbereite[n]' (§13 Abs. 1 TMG), muss eine Unterrichtung vor deren Einsatz erfolgen. Automatisierte Verfahren ist ein Oberbegriff für Technologien, die ohne das direkte Zutun des Nutzers Daten erheben. Ausprägungen können Cookies sein. Das sind kleine Dateien, die beim Nutzer gespeichert werden und deren Inhalt bei Anfragen an den Anbieter mit übertragen werden. Alternativen sind transpa-

rente, nur ein Pixel große Bilder, bei deren Abruf der Anbieter die IP-Adresse des Nutzers speichert, oder Web-Statistikwerkzeuge wie Google Analytics[5]. Daten, die über automatisierte Verfahren erhoben werden, können unter Umständen zu umfassenden Profilen über das Surf- oder (Online-) Einkaufverhalten der Nutzer führen.

Die oben beschrieben Pflichten kann ein Nutzer aufgrund ihre Sichtbarkeit von außen selbst auf Einhaltung überprüfen, beispielsweise durch das Analysieren der Datenschutzerklärung. Dies ist bei den nun folgenden Pflichten schwierig bis nicht möglich.

Auskunftspflicht Das TMG gibt dem Nutzer in §13 Abs. 7 TMG das Recht auf Auskunft. Das heißt, der Anbieter ist verpflichtet dem Nutzer Auskunft über seine personenbezogenen Daten oder über die zu einem Pseudonym gespeicherten Daten zu erteilen. Das Recht auf Auskunft ist die Grundlage für weitere Rechte, zum Beispiel das auf Sperrung, Berichtigung und Löschung [KSS08, Seite 267]. Der Anbieter kann die Auskunft auf Verlangen des Nutzers auch elektronisch erteilen.

Anonyme und pseudonyme Nutzung und Bezahlung Gemäß §13 Abs. 6 hat der Diensteanbieter die 'Nutzung von Telemedien und ihre Bezahlung anonym oder unter Pseudonym zu ermöglichen, soweit dies technisch möglich und zumutbar ist'. Wie genau eine Anonymisierung technisch erfolgen kann, beschreiben wir in Abschnitt 2.4.2. Anonymisierung dient dem Grundsatz der Datenvermeidung (§3 BDSG). Während für viele Dienste eine pseudonyme oder anonyme Nutzung möglich und auch zumutbar ist, beispielsweise beim Stöbern in dem Angebotskatalog eines Online-Shops, ist die anonyme oder pseudonyme Bezahlung deutlich schwieriger zu erreichen. Eine Möglichkeit sind Bezahldienste wie PayPal[6]. Einmal bei PayPal registriert, erfolgt die Abrechnung mit Diensten immer über PayPal, ohne dass der Nutzer alle personenbezogenen Daten auch nochmal bei jedem genutzten Dienst hinterlegen muss. Es sei jedoch darauf hingewiesen, dass die anonyme oder pseudonyme Bezahlung dazu führen kann, dass der Bezahldienst ein umso umfangreicheres Profil über die Einkäufe bei gleich mehreren Shops gleichzeitig erstellen kann.

Sicherstellung des Systemdatenschutzes 'Der Diensteanbieter hat durch technische und organisatorische Vorkehrungen sicherzustellen' (§13 Abs. 4 TMG), (i) dass der Nutzer die Dienstnutzung jederzeit beenden kann. Ebenso hat er sicherzustellen, (ii) dass personenbezogene Daten über den Zugriff auf den Telemediendienst oder dessen Nutzung unmittelbar nach der Beendigung des Dienstes gelöscht oder gesperrt werden, (iii) dass der Nutzer den Telemediendienst gegen die Kenntnisnahme Dritter geschützt

[5] `http://www.google.com/intl/de/analytics/` zuletzt gesichtet Juni 2010
[6] `www.paypal.de`

nutzen kann und dass (iv) personenbezogene Daten über die Nutzung verschiedener Telemedien durch denselben Nutzer getrennt verwendet werden. Weiter (v) dürfen Daten nur für Abrechnungszwecke zusammengeführt und (vi) Nutzungsprofile nicht mit Angaben zur Identifikation des Trägers eines Pseudonyms korreliert werden. Der Systemdatenschutz etabliert also unter anderem 'Datenschutz durch Technik' [KSS08, Seite 269].

2.2.4. Auslegung von Normen

Das Datenschutzrecht in Deutschland ist generisch in dem Sinne, als dass es auf viele existierende, aber auch noch in der Zukunft neu entstehende Technologien angewendet werden kann. Um zu überprüfen, nach welchem Gesetz ein Rechtsverhältnis besteht, beziehungsweise wie eine Norm zu verstehen ist, bedarf es oftmals der Auslegung. Wir werden auf diesen Abschnitt verweisen, wenn wir in Kapitel 5.1 und Kapitel 5.2 untersuchen, welche Verstöße nach aktuellem Datenschutzrecht bei unterschiedlichen Diensteanbietern im Internet vorliegen. Die Grundregeln der Auslegung gehen auf Friedrich Carl von Savigny [Sav40, Seite 213ff] zurück. Auslegung geschieht, indem man sich auf den Standpunkt des Gesetzgebers stellt 'und dessen Tätigkeit in sich künstlich wiederholt, also das Gesetzt in [...] sich von Neuem entstehen lässt' [Sav40]. Savigny unterscheidet vier Arten der Auslegung: die grammatische, (teleo)logische, historische und systematische Auslegung.

Grammatische Auslegung Die grammatische Auslegung bezieht sich auf das Wort, welches den Übergang des Denken des Gesetzgebers in unser Denken vermittelt. Der Sinn einer Rechtsnorm soll sich dabei möglichst nahe am Sinn des Wortes beziehungsweise der verwendeten Fachsprache orientieren.

(Teleo)Logische Auslegung Die logische Auslegung bezieht sich auf die Gliederung des Gedankens, also auf das logische Verhältnis, in welchem die einzelnen Teile sinn- und zweckgemäß zueinander stehen.

Historische Auslegung Die historische Auslegung berücksichtigt die Gegebenheiten, in die die Rechtsregeln zum Zeitpunkt ihrer Erstellung eingreifen sollten.

Systematische Auslegung Die systematische Auslegung berücksichtigt, in welchem Verhältnis eine betrachtete Rechtsregel oder ein Gesetz zum gesamten Rechtssystem steht und in welcher Art und Weise es in das System eingreifen sollte.

Es existieren weitere Methoden der Auslegung, die für diese Arbeit jedoch nicht relevant sind. Wesentlich hier ist, dass es trotz der generischen Natur des Datenschutzrechts eine Methodik gibt, mit Hilfe derer eine Norm gegen einen konkreten Sachverhalt überprüft werden kann.

2.3. Studien zur Privatheit

2.3.1. Nutzerstudien

In diesem Abschnitt beschreiben wir Nutzerstudien, zuerst allgemein zu Privatheit und anschließend speziell für kollaborative Suchmaschinen und standortbezogene Dienste. Wir weisen jeweils den Zusammenhang zu dieser Arbeit aus.

Internet-Nutzerstudien Nutzerstudien haben gezeigt, dass etwas 90% aller Internetnutzer besorgt um ihre Privatheit sind [AG05]. Gleichzeitig zeigen [OGH05, AG05], dass Nutzer gewillt sind, sensible, personenbezoge Daten auszutauschen, wenn angemessene Mechanismen zum Schutz der Privatheit vorliegen.

Wir werden in dieser Arbeit Anforderungen erarbeiten, wann solche Mechanismen 'angemessenen' sind. Dazu müssen zuerst die Privatheitspräferenzen der Nutzer identifiziert werden, beispielsweise wie in Kapitel 4.

Dies ist jedoch schwierig. [Ack00] zeigt, dass es einen Unterschied zwischen dem gibt, was ein Anbieter aus Sicht der Nutzer an Privatheitsmechanismen anbieten muss und dem, was aus technischer Sicht machbar ist. Ein besonderes Problem stellt sich außerdem in Hinblick auf die Methodik: [SGB01] zeigt, dass Präferenzen, die nur in Form von Fragebögen erhoben werden, nicht die wahren Präferenzen der Nutzer widerspiegeln. Wir analysieren aus diesem Grund in unseren Studien das Verhalten von Nutzern anhand konkreter Technologien, und wir trainieren unsere Nutzer auf den Implementierungen.

Nutzerstudien kollaborativer Suchmaschinen Kollaborative Suchmaschinen sind eine neue Technologie, zu der aktuell nur wenige Nutzerstudien verfügbar sind.

[OGH05] zeigt, dass Nutzer gewillt sind, Informationen auszutauschen. Bei CSE unterscheiden die Nutzer dabei insbesondere zwischen den Personengruppen Freunde, Verwandte und Kollegen [MH07]. Die Studien [LBR02, Mor07, TNP97] zeigen, dass Nutzer bereits heute für komplexe Suchprojekte, wie Urlaubsplanung oder Hausaufgaben, kollaborieren. [Mor07] zeigt, dass mehr als 85 Prozent der 'relativ erfahrenen Internetsucher' ihrer Studie Suchergebnisse austauschen. 25 Prozent der Befragten tun dies wöchentlich, 75 Prozent monatlich. Der Austausch findet dabei zumeist über E-Mail oder einen Messenger-Dienst statt. Die Nutzerstudie [MH07] zeigt, dass kollaborative Suchmaschinen die Effizienz des Austausches steigern. Eine Nutzerstudie der CSE 'MUSE' hat gezeigt, dass die Kommunikation unter den Nutzern bei der Kollaboration von entscheidender Bedeutung ist [RJK08].

Trotz dass die beschrieben Studien auf eine große Bereitschaft der Nutzer hindeutet, oftmals auch sensible Informationen wie Anfragen auszutauschen, gibt es keine Untersuchungen, die darauf abzielen, die Privatheitspräferenzen der Nutzer explizit zu analysieren. Mit unserer Studie versuchen wir diese Lücke zu schließen.

Nutzerstudien standortbezogener Dienste Studien zu standortbezogenen Diensten, wie die frühe Studie [BD03], untersuchen die Privatheitsauswirkungen verschiedener imaginärer standortbezogener Dienste. Sie befragen 16 Studienteilnehmer bezüglich des Nutzens verschiedener Dienste, wie oft sie die Dienste nutzen würden und welche Privatheitsbedrohungen sie befürchten. Das Hauptergebnis ist, das Dienste, die den Standort einer Person verfolgen, weit bedrohlicher sind als Dienste, die standortbewusst sind, beispielsweise abhängig von der Zeitzone.

In [LMD03] führen die Autoren eine Umfrage unter 130 Personen durch. Die Teilnehmer sollen ihre Privatheitspräferenzen für unterschiedliche Situationen angeben, beispielsweise 'Arbeit', 'Essen' oder 'Freizeitaktivitäten', und für unterschiedliche Personengruppen, die die Daten sehen könnten. Sie haben herausgefunden, dass die Präferenzen stärker von den Personengruppen abhängen als von der aktuellen Situation.

Zwei weitere Studien adressieren ein ähnliches Szenario: In [CSM+05] haben 16 Teilnehmer die Namen von realen Personen unterschiedlicher sozialer Gruppen genannt. Über zwei Wochen wurden den Teilnehmern anschließend zufällige, hypothetische Anfragen dieser Personen zugestellt. Anfragen betrafen die aktuelle Tätigkeit und was die Empfänger der anfragenden Person mitteilen würden. Im Ergebnis sind die Hauptursachen, ob eine Person Daten preisgeben würde oder nicht, die anfragende Person, der Grund der Anfrage und ob die Antwort auf die Anfrage für den Anfragesteller nützlich sein könnte. Im Gegensatz zu unserer Studie sind diese Anfragen aber alle hypothetischer Natur. Bei uns werden preisgegebene Informationen wirklich weitergegeben, und die Entscheidung etwas preiszugeben hat eine unmittelbare Auswirkung.

In [SC+05] haben die Autoren eine echte Applikation auf einem realen mobilen Endgerät eingesetzt. Das System verschickt Anfragen und automatische Mitteilungen immer dann an die Nutzer, wenn diese vorab definierte Regionen betreten. Die Positionsbestimmung erfolgt anhand der Mobilfunkwabe. Die Studie erfolgte mit 8 Teilnehmern (Entwickler der Applikation und deren Partner) über einen Zeitraum von fünf Tagen. Die kleine Menge der erhobenen Daten hat jedoch keine sinnvollen Schlussfolgerungen erlaubt. Die eingesetzte Applikation basierte auf SMS-Kommunikation. In unserer Studie sind die Teilnehmer permanent verbunden und tauschen ihre Position als auch Inhalte kontinuierlich aus. Wir bieten außerdem eine deutlich höhere Genauigkeit bei der Positionsbestimmung durch GPS, und die Bereiche, die Nutzer bei uns definieren, können sehr exakt bestimmt werden.

2.3.2. Anbieterstudien

Während wir zuvor Nutzerstudien betrachtet haben, beschreiben wir in diesem Abschnitt ausgewählte Studien zu Anbietern.

Soziale Netzwerkseiten sind die Klasse von Diensten, über die im Kontext Privatheit am meisten diskutiert wird. Das liegt insbesondere an der Größe der Nutzergruppe von mehreren (hundert) Millionen Personen, als auch daran, dass die Nutzer die zu schüt-

Tabelle 2.1.: Pflichtangaben zur Registrierung bei sozialen Netzwerkseiten

	VZ-Netzwerke	Xing	Facebook	MySpace
Vorname	✓	✓	✓	✓
Nachname	✓	✓	✓	✓
Geburtstag	✓		✓	✓
Geschlecht	✓	✓	✓	✓
Land/Region/Ort	✓	✓		✓
E-Mail	✓	✓	✓	✓
Status		✓		
Schule	✓			

zenden Daten selbst preisgeben. Tabelle 2.1 gibt eine Übersicht der Pflichtfelder bei vier ausgewählten sozialen Netzwerkseiten. Oftmals geben Nutzer in den Profilen weitere sensible Informationen an, wie Ansichten zu Religionen, welche Freundschaftsbeziehungen bestehen etc. Schutzmechanismen in sozialen Netzwerkseiten sind bisher rudimentärer Natur. Im Allgemeinen kann ein Nutzer über Häkchen bestimmen, welche Gruppe von Personen welche Inhalte sehen darf. Die Gruppen sind zumeist Freunde, Freundesfreunde, Personen einer bestimmten Schule oder aus einem bestimmten Ort, am Netzwerk angemeldete Nutzer und 'Jeder'. Inhalte sind Angaben zur Registrierung, Beiträge auf der Pinnwand, Nachrichten, Sichtbarkeit von Kontakten, Bilder, Verlinkungen von Bildern und Profilen, Kommentare und vieles mehr. [sti10] hat soziale Netzwerkseiten auf deren Datenschutzpraktiken hin untersucht. Zusammengefasst: Alle haben Mängel. Die VZ-Netzwerke schneiden am besten ab, LinkedIn und MySpace am schlechtesten. Die Mängel liegen dabei insbesondere bei der Transparenz, dem Umgang mit den Nutzerdaten, der Datensicherheit, der Möglichkeit, Rechte für Nutzer zu vergeben und dem Nutzerschutz.

[Xam08] hat die formularmäßige Erhebung personenbezogener Daten in Kontaktformularen analysiert. 42 Prozent der untersuchten Unternehmen setzen Kontaktformulare ein. 17 Prozent informieren über die Datenerhebung, nur 5% weisen einen direkten Link auf die Datenschutzerklärung aus. In Summe liegt bei 35 Prozent der Anbieter ein Verstoß bei der Erhebung personenbezogener Daten vor.

[Xam09] untersucht den Einsatz von Google AdSense anhand von 24.376 Webpräsenzen. Wenn die Zahl auch niedrig ist, so setzen 1,3 Prozent der Anbieter Google AdSense ein. 32 Prozent dieser Anbieter weisen auf den Einsatz hin, das heißt, 68 Prozent setzen sich über die Anforderung von Google hinweg, auf den Einsatz hinzuweisen. Außerdem hat [Xam09] den Einsatz von Google Analytics untersucht. Das Unabhängigen Landeszentrum stufte 2009 die Nutzung von Google Analytics durch Webseitenbetreiber als nicht mit dem deutschen Datenschutzrecht vereinbar ein [Una09]. 13 Prozent der untersuchten Anbieter nutzen trotzdem Google Analytics, 4 Prozent einen

vergleichbaren Dienst. 65 Prozent der Webpräsenzen setzen Google Analytics heimlich ein.

Die beschriebenen Studien analysieren wichtige Aspekte beim Umgang von Anbietern mit sensiblen, personenbezogenen Daten. Während aus wissenschaftlicher Sicht die isolierte Betrachtung der Aspekte sinnvoll sein kann, hilft diese den Nutzern nur bedingt weiter. Ein Nutzer kann nicht viele verschiedene, schnell vergängliche Statistiken konsultieren bevor er sich bei einem Anbieter anmeldet. Außerdem sind viele Verstöße nicht automatisiert zu erkennen. In Kapitel 5.2 gehen wir einen Mittelweg: Wir erheben, wo es möglich ist, Informationen über die Datenschutzpraktiken von Unternehmen automatisiert, führen diese in einem System zusammen, lassen aber Nutzer kollaborativ die Aspekte bewerten, bei denen das automatisiert nicht möglich ist.

2.4. Technische Ansätze zum Schutz der Privatheit

In Abschnitt 2.2.1 haben wir den Begriff des Systemdatenschutzes eingeführt. Neben organisatorischen Vorkehrungen zum Schutz der Privatheit ist die zweite Säule des Systemadatenschutzes der Schutz der Privatheit durch Technik.

In diesem Abschnitt stellen wir solche technischen Ansätze zum Schutz der Privatheit vor, sogenannte 'Privacy-Enhancing Technologies' (PETs). Eine vollständige Betrachtung ist aufgrund des Umfangs nicht möglich. Stattdessen betrachten wir die PETs, die charakteristische für solche sind, die wir in Kapitel 4 untersuchen und relevant sind in Hinblick auf unser PET in Kapitel 5.2. In einem eigenen Absatz stellen wir Anonymisierungstechniken vor, sowohl für Suchprotokolle als auch für standortbezogene Dienste.

Vorwegnehmend kann man sagen, dass es nicht den universellen Mechanismus für alle Privatheitsprobleme gibt. Vielmehr hängt der Mechanismus von drei zentralen Merkmalen ab: *erstens* von der betrachteten Technologie, zum Beispiel soziale Netzwerkseiten, standortbezogene Dienste oder Suchmaschinen. *Zweitens* gilt es zu unterscheiden, was eine Person wann vor wem schützen möchte. Eine Person könnte alle Daten oder nur Auszüge, vor dem Anbieter oder nur vor bestimmten Personen, zu jeder oder nur einer bestimmten Zeit schützen wollen. Kurz, eine in einem Szenario gewünschte Privatheitspräferenz kann in einem anderen Szenario inakzeptabel sein [WLW98]. *Drittens* ist für die Nutzung entscheidend, ob der Nutzer zum Schutz seiner Privatheit aktiv eingreifen muss oder ob der Mechanismus automatisiert im Hintergrund arbeitet.

2.4.1. PETs für das Web 2.0

PETs für kollaborative Suchmaschinen Wenige der klassischen Suchmaschinen unterstützen Privatheitsmechanismen. Der AskErasor von Ask.com erlaubt Nutzern die Löschung der Historie vergangener Anfragen und das Abschalten jeglicher Personalisierungsfunktionen. Auch Google bietet mittlerweile an, die Suchhistorie abzuschalten.

Das führt zu einer Abwägung: Die Personalisierung einfach abzuschalten erhöht die Privatheit. Gleichzeitig senkt dies aber die Effektivität, das heißt die Qualität des Suchergebnisses [Gla01, KF07, TAJP07, WM07] und die Einsatzmöglichkeit von zum Beispiel Werbung (Kapitel 6). Gerade Werbung ist aber der Hauptgrund, warum viele Dienste den Nutzern kostenlos angeboten werden können.

Die kollaborativen Suchmaschinen SearchTogether [MH07], MUSE [RJK08] und I-SPY [SBB+05] lassen, anhand von Suchsitzungen oder Nutzergruppen, die Nutzer manuell entscheiden, welche Information mit wem ausgetauscht werden soll. Das lässt die Nutzer zwar theoretisch ihre Präferenzen durchsetzen, wir werden in Kapitel 4 jedoch zeigen, dass Mechanismen, die eine kontinuierliche Aufmerksamkeit von einem Nutzer erfordern, versagen. Erlauben Nutzer beispielsweise bei den genannten CSEs anderen Nutzern Zugriff auf eine Suchsitzung, so bekommen diese nicht nur zukünftige Anfragen angezeigt, sondern die gesamte Historie. Es ist anzunehmen, das mangelnde Aufmerksamkeit hier schnell zu einem Privatheitsproblem führen wird.

PETs für standortbezogene Dienste Privatheitsmechanismen für standortbezogene Dienste sind unterschiedlich komplex. Der einfachste Mechanismus ist ein Schaltern zum (De-) Aktivieren des GPS [BD03]. Komplexer sind PETs, bei denen Nutzer zwischen unterschiedlichen sozialen Gruppen unterscheiden und die Akkuratheit jeder einzelnen preisgegebenen Information variieren können [CSM+05]. Automatische Mechanismen umfassen [GG03, GL04, KYS05], die wir im Kontext von Anonymisierung (Abschnitt 2.4.2.4) untersuchen. Wir berücksichtigen in unserer Studie (Kapitel 4.2), dass Mechanismen unterschiedlich komplex sind und untersuchen, inwieweit Nutzer den PETs vertrauen und mit welchen PETs sie ihre Privatheitspräferenzen durchsetzen können.

Kollaborative Ansätze zum Schutz der Privatheit Es existiert eine Vielzahl von Untersuchungen zu Datenschutzproblemen bei kollaborativen Systemen [Fel07, FE08, NS09]. Weit weniger Ansätze nutzen jedoch das Wissen und die Erfahrungen einer Gemeinschaft von Nutzern gerade zum Schutz der Privatheit. Wir werden solch einen Ansatz in Kapitel 5.2 vorstellen. Im Folgenden beschreiben wir einige verwandte Arbeiten im Detail. Diese basieren insbesondere auf dem Einsatz von Tagging, das heißt der Verschlagwortung von Objekten und Orten und aus daraus erstellten Folksonomien. Folksonomien sind Kollektionen von Annotationen, erstellt durch eine Vielzahl

von Nutzern. Eine exakte Definition des Begriffs Folksonomie findet sich in [SSP09, Mat04].

[BBR08] schlägt eine Web 2.0-Architektur (Privacy2.0) vor, bei der eine Gemeinschaft von Nutzern eine Folksonomie potentieller Datenschutzprobleme erstellt. Nutzer taggen Privatheitsbedrohungen, beispielsweise kritische Webseiten, oder mit Hilfe mobiler Endgeräte Kameras an öffentlichen Plätzen, oder Geschäfte, die intensiv RFID einsetzen. Webseiten können dabei für einen Anbieter stehen, aber auch für zum Beispiel die Profilseite anderer Nutzer in sozialen Netzwerkseiten. Über die vorgeschlagene Architektur werden diese Informationen zusammengeführt. Surft ein andere Nutzer nun eine bereits annotierte Webseite an, wird er über die in Form von Tags hinterlegten Erfahrungen anderen Nutzer informiert und gegebenenfalls gewarnt. Eine Implementierung von Privacy2.0, integriert in den Browser Firefox, findet sich in [HBB10] unter dem Namen 'Taxor'. Anhand einer Nutzerstudie haben die Autoren erste Erkenntnisse gewonnen, dass ein System wie Privacy2.0 tatsächlich helfen kann, Privatheitsbedrohungen zu identifizieren.

[8] schlägt einen Ansatz 'Primo' vor, der auf semantischen Annotationen von Fotos basiert. Die Nutzer kollaborieren, indem sie Zuordnungen von Fotos zu den abgelichteten Personen vornehmen. Primo lernt diese Zuordnungen und informiert für die Zukunft ihm bekannte Personen, wenn über sie Fotos ins Internet und insbesondere auf soziale Netzwerkseiten gestellt werden. Außerdem erlaubt Primo die Erstellung von Regeln, die definieren, wer welche Fotos wann sehen darf.

[SSP09] betrachtet das Problem der Teilhaberschaft (Co-Ownership) an Inhalten, die Nutzer auf soziale Netzwerkseiten hochladen. Das prominenteste Beispiel für Inhalte sind Bilder. Ein Bild kann einen Ersteller haben, jemand anderes kann dieses Bild in das Internet einstellen, und mehrere Personen können auf einem Bild abgelichtet sein. Die Autoren schlagen einen auf Spieltheorie basierenden Ansatz vor, um die Privatheitspräferenzen aller beteiligten Parteien, der sogenannten Besitzer, zu berücksichtigen. Der auf dem Modell 'Clarke-Tax' [Cla71] beruhende Ansatz erlaubt es den Parteien, die die Inhalte hochladen, die Teilhaberschaft an den Inhalten an andere Nutzer weiterzugeben. Gemäß einer Auszahlungsfunktion bekommt die die Teilhaberschaft weitergebende Partei, als auch die Partei, die die Teilhaberschaft erhält, vom System Punkte. Um eine Einigung zwischen allen Teilhabern auf eine Privatheitspräferenz zu erzielen, können die Nutzer diese Punkte einsetzen und über einen Auktionsmechanismus auf die Präferenz bieten, die sie für das betrachtete Datenobjekt durchsetzen möchten. Das verfügbare Budget jeden Nutzers setzt sich aus den beschriebenen Auszahlungen zusammen. Ohne näher auf 'Clark-Tax' einzugehen garantiert dieser Ansatz eine relativ faire Verteilung der Auszahlung (Nutzen), das Prinzip ist einfach und es ist nicht manipulierbar. Da die Definition von Privatheitspräferenzen für viele Datenobjekte, die Weitergabe der Teilhaberschaft und das Bieten auf die Präferenzen aufwändig sind, schlagen [SSP09] in einem zweiten Schritt vor, Annotationen (Tags) der Bilder zu nutzen, um ähnliche Bilder zu identifizieren. Auf ähnliche Bilder wird dann die gleiche

Privatheitspräferenz angewendet. Der Ansatz wurde beispielhaft in Facebook implementiert. Offen bleibt die Frage, warum ein Nutzer, der einen Inhalt hochlädt und seine Datenschutzpräferenz durchsetzen möchte, Teilhaberschaften weitergeben sollte. Auch wird nicht klar, wie ein Schutz entstehen soll, wenn viele Bilder einer kompromittierenden Situation online gestellt werden. Möchte nur eine Person nicht, dass ein Bild online gestellt wird, viele andere hingegen schon, wird der betroffenen Person irgendwann das Budget zum Bieten ausgehen. Das ist problematisch, da gegebenenfalls ein einzelnes Bild gleichermaßen die Privatheit verletzt, wie unter Umständen viele Bilder der gleichen Situation.

[V+09] basiert ebenfalls auf Annotationen. Die Autoren konzentrieren sich auf das Finden ähnlicher Privatheitspräferenzen. Konkreter nutzen sie die Tags, um Nutzer ähnlicher Privatheitspräferenzen zu clustern. Sind Nutzer mit ähnlichen Präferenzen identifiziert, werden für einen Nutzer, der für eine Webseite noch keine Präferenz formuliert hat, die Präferenzen der anderen Nutzer aus dem gleichen Cluster herangezogen.

Tagging ist einfach und intuitiv, bietet sich also an, wenn Nutzer ausdrücken sollen, was sie privat empfinden. Es birgt jedoch eine gewisse Unschärfe: Nutzer annotieren Objekte mit ihren eigenen Worten und in ihrer eigenen Sprache. Es braucht große Mengen von Daten, um Ähnlichkeiten zwischen Wortkonzepten und damit Nutzern feststellen zu können. Gleichzeitig ist aufgrund der Unschärfe nicht sichergestellt, dass Ansätze wie [V+09] die Präferenzen anwenden, die ein Nutzer auch wirklich wünscht. Im Kontext Privatheit ein ernstzunehmendes Problem, da einmal preisgegebene Daten für immer öffentlich zugänglich sein können. Ansätze wie [BBR08] betrachten außerdem nur subjektiv, was privat sein soll. Datenschutzrechtlich muss das keine Relevanz haben und ohne die Möglichkeit sein Recht einzuklagen, können Bedrohungen auch nur schlecht abgewendet werden. Wir stellen in Kapitel 5.2 einen Ansatz vor, der ebenfalls die Kollaboration von Nutzern unterstützt, Nicht-Experten die Nutzung unseres Systems erlaubt, Datenschutzverstöße aber systematisch und strukturiert anhand des Gesetzes identifiziert.

In Kapitel 5.1 stellen wir eine umfangreiche Studie vor, bei der wir manuell Datenschutzpraktiken von Anbietern untersucht haben. Wir beschreiben im Folgenden, warum existierende PETs Datenschutzverstöße nicht automatisiert erkennen können. In Kapitel 5.2 stellen wir selbst einen Ansatz vor, wie Nutzer Datenschutzverstöße von Anbietern identifizieren können. Hier beschreiben wir, warum es erforderlich ist, einen neuen Weg zu gehen und nicht auf einer existierenden Technologie aufzusetzen.

Plattform for Privacy Preferences Project (P3P) Die ‘Plattform for Privacy Preferences’ (P3P) ist ein vom World Wide Web Consortium (W3C) standardisiertes Protokoll [Mar02]. Es ist der wahrscheinlich bekannteste Mechanismus, um Nutzer vor

Anbietern zu warnen, deren Datenschutzpraktik in Konflikt mit den Datenschutzpräferenzen des Nutzers steht. P3P ermöglicht es Anbietern von Webseiten, die Datenschutzerklärung strukturiert (XML) und damit automatisiert verarbeitbar zu formulieren.

Dem Nutzer stehen verschiedene Sprachen zur Verfügung, um sein Präferenzen zu formulieren [CLM, Hog02, AKSX03]. Agenten, zum Beispiel kleine Programme integriert in den Browser, gleichen die vordefinierte Privatheitspräferenzen mit den P3P-Datenschutzerklärungen, das heißt den Datenschutzpraktiken, ab. [CGA06] gibt einen Überblick über existierende Agenten. Die bekannteste Implementierung ist PrivacyBird[7]. Im Falle eines Konfliktes zwischen der Datenschutzpraktik des Anbieters und der Privatheitspräferenz des Nutzers kann der Agent den Nutzer warnen oder gar den Zugriff auf die Seite sperren. Ein solcher Fall liegt zum Beispiel vor, wenn der Anbieter Daten weitergeben möchte, der Nutzer dies aber ablehnt.

Der Ansatz von P3P klingt vielversprechend, es gibt jedoch zwei wesentliche Nachteile: Der *erste* Nachteil ist die mangelhafte Ausdrucksmächtigkeit von P3P. Ein Standard wie P3P kann nicht alle möglichen Sachverhalte einer in natürlicher Sprache verfassten Datenschutzerklärung abdecken [RDM09]. Dies gilt beispielsweise in Bezug auf die EU-Gesetzgebung [JZ05]. So kann die Weitergabe personenbezogener Daten in Sonderfällen, zum Beispiel wenn das Leben einer Person auf dem Spiel steht, nicht ausgedrückt werden. Dies trifft aber auch auf Fälle der weltweiten Datenweitergabe zu. Gerade die Verfolgung personenbezogener Daten beim internationalen Austausch ist aber ein zentrales Problem. Da sich die Datenschutzgesetzgebung international im permanenten Wandel befindet, ist auch nicht davon auszugehen, dass P3P jemals in der Lage sein wird, alle Eventualitäten abzudecken. *Zweitens* hat sich P3P nicht durchgesetzt. So nutzen nach den Studien [BRDM07, RDM09] nur 3,5 Prozent der großen Anbieter P3P. Darüber hinaus weisen die P3P-Datenschutzerklärungen Unterschiede zu den Erklärungen im Klartext auf und bis zu 75 Prozent der P3P-Datenschutzerklärungen sind auf technischer Ebene fehlerhaft [ECC06]. Das heißt, sie entsprechen nicht der vorgegebenen Struktur und können somit gegebenenfalls nicht ausgewertet werden. Verlässt sich ein Nutzer auf die P3P-Erklärung, wähnt er gegebenenfalls irrtümlich seine Präferenzen als berücksichtigt, was die Gefahr für seine Privatheit unter Umständen verstärkt.

Verarbeitung natürlicher Sprache Aus oben genannten Gründen haben die Autoren von [BKK06] versucht, die natürlichsprachigen Datenschutzerklärungen auszuwerten. Dazu bedienen sie sich Techniken der Verarbeitung natürlicher Sprache (engl. Natural Language Processing, NLP). Während im Vergleich zu P3P das Problem bleibt, dass nur das erkannt werden kann, was die Entwickler von [BKK06] auch vorher modelliert haben, entsteht zusätzlich das Problem, dass die gleichen Sachverhalte auf unterschiedlichste Weise formuliert sein können. Das wiederum erfordert eine aufwändi-

[7] AT&T, PrivacyBird, `http://www.privacybird.com`, Mai 2010

ge, von Menschen durchzuführende Vorverarbeitung der Datenschutzerklärung, damit diese von dem in [BKK06] verwendeten Regelinterpretierer verstanden wird. Bei der riesigen Anzahl von Anbietern im Internet (und dem heutigen Stand der Forschung) halten wir diesen Ansatz für nicht praktikabel.

Generatoren von Datenschutzerklärungen Die beschriebene NLP Technik erfordert es, die Datenschutzerklärungen der Anbieter vorzuverarbeiten, das heißt gemäß eines vorgegebenen Formates umzustrukturieren. Generatoren von Datenschutzerklärungen arbeiten umgekehrt und stellen Textbausteine zur Verfügung, die zu einer Datenschutzerklärung kombiniert werden können. Beispiel für solche Generatoren sind der OECD Privacy Statement Generator[8] und der Privacy Policy Generator[9]. Unter der Annahme, dass jedes Unternehmen seine Datenschutzerklärung mit solch einem Werkzeug generiert, würde die Erklärungen sicherlich vereinheitlicht und somit die Verständlichkeit verbessert. Das Hauptproblem liegt jedoch darin, dass diese Tools nur formulieren, was man ihnen vorgibt, nicht was die Anbieter tatsächlich praktizieren. Beispielsweise steht der Einsatz von Cookies und die Aussage 'kein Einsatz von Verfahren zur automatisierten Verarbeitung personenbezogener Daten' in Konflikt.

Zusammenfassend kann gesagt werden, dass es eines Mechanismusses bedarf, der sowohl berücksichtigt, was ein Anbieter als seine Praktik ausweist, als auch welche Praktik er tatsächlich einsetzt. Ein Mechanismus muss mehr als die Datenschutzerklärung berücksichtigen, beispielsweise auch den Registrierungsprozess, den Prozess der Einwilligung, den Einsatz von Web-Statistikwerkzeugen etc. Er muss leicht an eine sich verändernde Gesetzgebung anpassbar sein und ein Nutzer muss den Mechanismus auch ohne Zutun des Anbieters einsetzen können, beispielsweise ohne das freiwillige Bereitstellen einer P3P Datenschutzerklärung.

2.4.2. Anonymisierung

Ein zentrales Prinzip zum Schutz personenbezogener Daten ist Anonymisierung. Dieser Aspekt ist wichtig für diese Arbeit: Nutzer der kollaborativen Suchmaschine fordern Anonymität (Kapitel 4.1), wir testen ein auf Anonymisierung basierendes PET für standortbezogene Dienste (Kapitel 4.2), und wir anonymisieren Suchprotokolle (Kapitel 6). Ein Datenbestand ist anonym, wenn keine Person eindeutig identifiziert werden kann [Par95] (siehe auch BDSG §3 Abs. 6). Während bei der Pseudonymisierung die Identifikationsmerkmale durch ein Pseudonym ersetzt werden, ist die Idee der Anonymisierung der Schutz vor der Verknüpfung korrelierenden Wissens.

[8] `http://www.oecd.org:80/sti/privacygenerator` , Mai 2010
[9] `http://policygenerator.net`

Es gibt unterschiedliche Definitionen von Anonymität. Zu jeder dieser Definitionen gibt es im Allgemeinen unterschiedliche Methoden, den Datenbestand zu manipulieren, bis er der Definition genügt, das heißt anonym ist. Wir werden die Definitionen in Abschnitt 2.4.2.1 vorstellen. Methoden zur Manipulation des Datenbestandes beschreibt Abschnitt 2.4.2.2.

2.4.2.1. Definitionen von Anonymität

Im Folgenden stellen wir die prominentesten Definitionen von Anonymität vor und insbesondere, wie diese aufeinander aufbauen.

Trennung von sensiblen und identifizierenden Attributen Die Mehrheit der Definitionen beruht auf der Bildung von Äquivalenzklassen identifizierender Attribute. Das bedeutet entsprechend, dass ein Datensatz zu einer Person in identifizierende und sensible Attribute unterteilt werden muss. Im Kontext von Gesundheitsdaten könnte beispielsweise die Kombination Postleitzahl, Geburtsdatum und Geschlecht identifizierend sein, das heißt einen Quasi-Identifikator bilden. Die Art der Erkrankung stellt die schützenswerte sensible Information dar. Der identifizierende Teil des Datensatzes wird anschließend solange verändert, bis er identisch zu einer vorgegebene Mindestanzahl von identifizierenden Teilen anderer Datensätzen und daher nicht von diesen zu unterscheiden ist.

[Swe02] hat diesen Ansatz unter dem Namen *k-Anonymität* wesentlich geprägt. k-Anonymität schützt mit einer Konfidenz von 1/k vor einer 'korrekten' Verknüpfung korrelierenden Wissens. Mit anderen Worten muss jede Äquivalenzklasse mindestens k Datensätze repräsentieren. Noch anders ausgedrückt ist jeder identifizierende Teil eines Datensatzes von mindestens k-1 identifizierenden Teilen anderer Datensätze nicht unterscheidbar.

Ein Problem mit der Idee der k-Anonymität liegt in der Möglichkeit, dass die sensiblen Teile aller Tupel, die in dieselbe Äquivalenzklasse fallen, identisch sind (Homogenitätsproblem).

Beispiel 2: Angenommen man weiß, dass die Krankenakte einer bestimmten Person aus einer bestimmten Stadt in einem Krankenhaus sein muss. Fallen nun zufällig alle Personen aus der gleichen Stadt in eine gemeinsame Äquivalenzklasse, so ist, sollten alle Personen aus dieser Stadt die gleiche Krankheit haben, diese Krankheit direkt ablesbar. Das heißt, die Privatheit ist verletzt.

Die auf k-Anonymität aufbauende Definition von *l-Diversität* [MGKV06] erfordert, dass sich die sensiblen Attribute der Datensätze einer Äquivalenzklasse unterscheiden, beispielsweise in l unterschiedliche Krankheiten. So wird dem in Beispiel 2 beschriebenen Homogenitätsproblem Rechnung getragen. Anders ausgedrückt wird, gegeben

korrelierendes Wissen, der Wissensgewinn zwischen 'vor der Preisgabe' eines Datenbestandes und 'nach der Preisgabe' eines Datenbestandes versucht klein zu halten.

l-Diversität garantiert zwar unterschiedliche sensible Attribute zu den Datensätzen einer Äquivalenzklasse, macht dies jedoch ohne Berücksichtigung statistischer Eigenschaften des Ausgangsdatenbestandes.

Beispiel 3: In einem Datenbestand sind 1 Prozent positive HIV-Befunde. Gefordert ist 2-Diversität gemäß obiger Definition, k sei 100. Kommen nun in einer Äquivalenzklasse 50 HIV Befunde vor, und weiß ein Betrachter, dass eine bestimmte Person in dieser Äquivalenzklasse liegt, muss er annehmen, dass die Person zu 50 Prozent HIV positiv ist. Das steht im Widerspruch zu einem Gesamtanteil von nur einem Prozent.

t-Closeness [LLV] garantiert nicht nur eine Diversität der sensiblen Attribute pro Äquivalenzklasse, sondern orientiert diese Diversität zusätzlich anhand der Verteilung der sensiblen Attribute in der Ausgangspopulation. Somit kann ein Angreifer aus der Gesamtverteilung eines sensiblen Attributes und der Verteilung des Attributes in der Äquivalenzklasse kein neues Wissen ableiten.

Das Verfahren *Anatomy* [XT06] trennt die sensiblen Attribute von den identifizierenden ab und speichert sie in getrennten Relationen. Gemäß den Definitionen von l-Diversität oder t-Closeness werden auch hier die identifizierenden Attribute in Äquivalenzklassen gruppiert. Die Autoren können jedoch zeigen, dass die Verschlechterung der statistischen Eigenschaften des anonymisierten Datensatzes durch Anatomy abgeschwächt werden.

Anonymisierung von Mengen sensibler Attribute Für viele Szenarien ist die Trennung von sensiblen und identifizierenden Attributen entweder nicht möglich oder nicht nützlich. Oftmals kann sowohl ein einzelnes Attribut, aber insbesondere auch deren Kombinationen die Privatheit gefährden. [TMK08] führt folgendes Beispiel an:

Beispiel 4: Man stelle sich eine Datenbank vor, in der zu mehreren Kunden gespeichert wird, welche Artikel diese jeweils eingekauft haben. Weiter habe ein Angreifer auf die Privatheit der Kunden Einblick in die Datenbank und Kenntnisse über eine Teilmenge der gekauften Artikel eines Kunden erlangt, beispielsweise weil die Artikel im Einkaufswagen obenauf gelegen sind. Hat der Angreifer beispielsweise gesehen, dass ein Kunde Glühbirnen und Milch gekauft hat, und hat sonst kein anderer Kunde diese beiden Artikel in Kombination gekauft, so kann der Angreifer alle Artikel, die der Kunde außer Glühbirnen und Milch in der gleichen Transaktion gekauft hat, dieser Person zuordnen.

Es ist hier also zur Erreichung von Anonymität nicht ausreichend, dass, wie wir wie in dem vorigen Absatz fordern, jeder Artikel von k-1 anderen Kunden ebenfalls gekauft wurde. Stattdessen muss eine geeignete Anonymisierung sicherstellen, dass aus dem anonymen Datensatz nicht ersichtlich ist, wer genau eine bestimmte Artikel-*Kombinationen* gekauft hat. Dabei kann man keine Vorhersage machen, welche Produkte in Kombination einem möglichen Angreifer bekannt sind. Ebenso ist unklar, wie

viele Produkte der Angreifer gesehen hat.

[TMK08] schlägt zum Schutz vor diesem Problem k^m-Anonymität vor.

Definition 1 *Ein Datensatz D ist annonym, wenn kein Angreifer Wissen über mehr als m Elemente einer Transaktion $t \in D$ hat, und mit Hilfe von m Elementen nicht weniger als k Datensätze aus D identifizieren kann.*

Bezogen auf Beispiel 4 müsste es gemäß der Definition von k^m-Anonymität also mindestens noch k weitere Kunden geben, die ebenfalls eine Glühbirne und Milch kaufen. Genauer muss jede Kombination aus m Elementen in k Transaktionen existieren.

Ein Problem bei dieser Definition von Anonymität ist die Wahl von m. Während die Autoren von [TMK08] kleine m wählen und je nach Privatheitspräferenz erhöhen, setzt [HN09] m auf die Größe der Transaktion, das heißt das maximale m. Interessanterweise können die Autoren von [HN09] zeigen, dass der mit der Anonymisierung einhergehende Informationsverlust bei sogar stärkerer Privatheit kleiner ist als in [TMK08].

2.4.2.2. Berechnung der Anonymisierung

Bisher haben wir unterschiedliche Definitionen von Anonymität kennengelernt. Damit ein Datenbestand einer Definition von Anonymität genügt, muss dieser im Regelfall verändert werden. Dazu gibt es vier Möglichkeiten: *Erstens* kann die anonymisierende Partei zum Erreichen der geforderten Anonymität Rauschen hinzufügen, das heißt Datensätze einfügen. *Zweitens* kann sie Informationen unterdrücken, das heißt Datensätze löschen. *Drittens* kann sie Daten vertauschen und *viertens* Datensätze generalisieren.

Der für diese Arbeit relevante Punkt ist das selbsterklärende Löschen von Daten und das Prinzip der Generalisierung. Generalisieren bedeutet, von einem konkreten Attribut eine verallgemeinerte Darstellung zu finden. Im Falle von kategorischen Attributen, zum Beispiel den Postleitzahlen `{68259, 68199, 68761}`, ist das die generalisierte Postleitzahl `68***`. So entsteht aus der Menge der exakten Postleitzahlen eine dreielementige Äquivalenzklasse. Binäre Attribute wie das Geschlecht (männlich / weiblich) können nur zu `*` generalisiert werden. Numerische Attribute werden bei der Generalisierung üblicherweise in Bereiche unterteilt, für das Attribut 'Alter' beispielsweise in 0-30, 30-60, 60-100 Jahre.

Nahezu jede der Arbeiten, die eine neue Definition von Privatheit vorstellt, bietet auch ein Verfahren an, einen Datensatz gemäß der Definition zu anonymisieren. Wir möchten an dieser Stelle Abstand davon nehmen, jedes der Verfahren vorzustellen. Wichtig zu wissen ist aber, dass bei der Anonymisierung eine Abwägung (Tradeoff) getroffen werden muss zwischen dem Grad der Anonymisierung, das heißt dem Schutz der Privatheit, und der Qualität des anonymisierten Datensatzes [Ada07]. Diese Qua-

lität umfasst im Allgemeinen statistische Eigenschaften, also beispielsweise wie viele Daten müssen für die Anonymisierung unterdrückt / generalisiert / eingefügt / verändert werden. Verfahren, die bezogen auf die Qualität optimal anonymisieren, sind für k-anonymitätsbasierte Definitionen NP-schwer [MW04]. Entsprechend sind viele der vorgeschlagenen Algorithmen Heuristiken, die in vertretbarer Zeit ein gutes Ergebnis, nicht zwangsläufig aber auch das optimale Ergebnis liefern.

2.4.2.3. Anonymisierung von Suchprotokollen

Einblicke in die Suchhistorie einer Person erlauben Rückschlüsse auf deren Interessen, Vorlieben, Gewohnheiten etc. Damit Suchanfragen einer Person zugeordnet werden können, müssen aus der Suchhistorie Identifikatoren extrahiert werden. Einfach ist dies durch das weit verbreitete Ego-Surfing, bei dem Personen ihr Selbstporträt im Internet mit Hilfe von Suchmaschinen prüfen [RRK05]. Dazu geben die Personen von sich aus den Quasi-Identifikator, beispielsweise den Namen und die Stadt in der sie wohnen, direkt in die Suchmaschinen ein. Speichert der Suchmaschinenanbieter diese Information als Historie ab, kann er relativ leicht Rückschlüsse auf den Urheber der Suchhistorie ziehen.

Aber auch ohne Ego-Surfing kann der Personenbezug hergestellt werden. [JKPT07] demonstriert einen Ansatz, wie Sequenzen von Anfragen Rückschlüsse auf das Geschlecht, Alter und Ort der anfragenden Person zulassen. Mit Hilfe eines Komitees von Klassifikatoren bilden die Autoren Gruppen von Nutzern mit gleichen Merkmalen.

Beispiel 5: Anhand eines Trainingsdatensatzes lernt der Ansatz, dass Wörter wie Braut, Make-up, Stricken, Haare, E-Cards, Funkeln, Yoga und Diät auf das weibliche Geschlecht hinweisen. Wörter wie NFL, Poker, ESPN, UFL, Autobahn, Prostata, Football, Golf, Männlich, Wrestling, und eine Vielzahl von Begriffen aus der Erotik, deuten auf Männer hin. Für Suchanfragen von Nutzern, die nicht Teil des Trainingsdatensatzes sind, entscheidet das Komitee, welches Geschlecht der betroffene Nutzer hat.

Die Autoren zeigen, dass die Akkuratheit für die Vorhersage des Geschlechts bei 83,4 Prozent liegt. Insbesondere zeigen sie, dass solch eine Zuordnung die Identifikation der Menge der möglichen Personen, die eine Anfrage abgesetzt haben können, um 300 bis 600 Prozent im Vergleich zu einer zufälligen Zuordnung verbessert.

Um solchen und ähnlichen Angriffen entgegenzuwirken gibt es Anonymisierungstechniken speziell für Suchhistorien. In [Ada07] schlagen die Autoren einen zweistufigen Prozess vor. Zuerst entfernen sie seltene Terme, da diese, im schlimmsten Fall nur einmal existierenden Terme, die Identifikation einer Person erlauben könnten. Seltene Terme wirken dabei als Quasi-Identifikatoren. Weiß ein Angreifer einen nur einmal vorkommenden Term einer Person aus dem Datensatz, so kann er auch alle anderen Suchterme dieser Person identifizieren.

Beispiel 6: Zur Illustration des ersten Schrittes nehmen wir an, dass die Terme 'Pasta'

von 100, 'WM-Ball' von 200 Personen und 'Reziprozität' von nur einer Person gesucht wurde. Alice hat nach allen drei Termen gesucht, repräsentiert durch die Tupel <ID_a, 'Pasta'> <ID_a, WM-Ball><ID_a, Reziprozität><ID_a lässt keine Rückschlüsse auf Alice zu. Weiß nun Bob, dass Alice nach 'Reziprozität' gesucht hat, kann Bob bei Zugriff auf den Datensatz den Schluss ziehen, dass Alice auch nach 'Pasta' und 'WM-Ball' gesucht hat. Entsprechend entfernt der erste Schritt in [Ada07] 'Reziprozität'.

Anschließend erstellt der vorgeschlagene Ansatz für jedes Thema, zu dem Suchanfragen abgesetzt wurden, einen eigenen virtuellen Nutzer. Danach ordnet der Algorithmus die Suchanfragen themenbezogen diesen virtuellen Nutzern zu.

Beispiel 7: Beispiel 6 fortsetzend bildet der zweite Schritt die Suchterme auf Themen ab. Alle drei Terme entstammen einem unterschiedlichen Kontext. Entsprechend werden drei virtuelle Nutzer angelegt, u_1 für 'Fußball', u_2 für 'Kochen' und u_3 für 'Spieltheorie'. Der Ansatz ordnet anschließend 'WM-Ball' dem Thema 'Fußball', 'Pasta' dem Thema 'Kochen' und Reziprozität dem Thema 'Spieltheorie' zu.

Der erste Schritt entfernt jedoch bereits 97 Prozent der unterschiedlichen Terme und 64 Prozent der Größe des Suchprotokolls. Darüberhinaus verringert die Zuordnung von Suchanfragen von einem realen zu mehreren virtuellen Nutzern den Nutzen des anonymisierten Logs noch weiter, zum Beispiel bei der Vorhersage der Eigenschaften eines durchschnittlichen Nutzers.

[KNPT07] analysiert einen anderen Weg namens 'token-based hashing' und zeigt dessen Ineffektivität als Anonymisierungstechnik. Bei diesem Ansatz wird eine Suchanfrage in sogenannte Token zerlegt. Anschließend bildet der Ansatz jedes dieser Token über eine Hashfunktion auf eine ID ab. Die Autoren prüfen, inwieweit es möglich ist, diese IDs wieder zurück auf die eigentlichen Terme abzubilden. Interessanterweise ist die Häufigkeit einer ID beziehungsweise eines Terms kein ausreichender Indikator. Die Kombination der Häufigkeit einer ID, die Häufigkeit einer ID in einer nur einmal vorkommenden Anfragen und das Vorkommen in Kombination mit anderen Termen hingegen schon. So erhalten sie eine Akkuratheit der Zuordnung der IDs zu den ursprünglichen Termen von über 94 Prozent.

[KKMN09] untersucht die Preisgabe von 'Anfrage-Klick-Graphen'. Dabei repräsentieren die Knoten sowohl Anfragen als auch URLs. Die Kanten zeigen, welche URLs Nutzer aufgrund welcher Anfrage angeklickt haben. [KKMN09] anonymisiert einen Anfrage-Klick-Graphen, indem sie (i) die Anzahl möglicher Anfragen einer Person limitieren, (ii) nur solche Anfragen preisgegeben werden, deren Häufigkeit plus ein über eine Zufallszahl abgeleitetes Rauschen größer einer Schranke ist, und (iii) immer 10 URLs aus dem Suchergebnis zu der preisgegebenen Anfrage veröffentlichen. Auf die Häufigkeit der Klicks pro URL wird ebenfalls ein Rauschen addiert. Mit diesem Ansatz und dem gegebenen Suchprotokoll können zwischen 27 und 34 Prozent aller Suchanfragen anonym preisgegeben werden, abhängig von dem verwendeten Grad der Anonymität.

Der dienstnehmerseitige Ansatz von [MC08b] versucht, durch das Einfügen von verschleiernden Anfragen (engl. cover queries) die Intention des Nutzers zu verbergen. Die Kunst hier ist es, gegeben eine Anfragen, eine Menge von Verschleierungsanfragen mit den gleichen Charakteristiken wie die der originalen Anfragen, aber mit einem anderen Informationsbedürfnis zu finden. So ist eine Anfrage gemäß der Definition der Autoren glaubhaft leugbar (engl. a plausible deniable search), wenn jede Anfrage aus der Menge der Verschleierungsanfragen und der Originalanfrage diese Menge mit gleicher Wahrscheinlichkeit erzeugen würde. Außerdem müssen alle Anfragen aus dieser Menge unterschiedliche Informationsbedürfnisse ausdrücken, und alle Anfragen müssen gleichermaßen glaubhaft sein. Obwohl die Evaluierung eher demonstrierenden Charakter hat, zeigt der Ansatz, dass es möglich ist, unabhängige und plausible Mengen von Anfragen zu generieren und zur Verschleierung eines Informationsbedürfnisses einzusetzen. Offen bleibt insbesondere, wie auch die Verwendung der Suchergebnisse verschleiert werden soll.

Vielversprechend für die Anonymisierung von Suchanfragen ist auch die bereits eingeführte Definition der k^m-Anonymität von [TMK08]. Anstelle von gekauften Produkten aus dem Kontext der Warenkorbanalyse (vgl. Beispiel 4) sind es die einzelnen Suchterme der abgesetzten Anfragen, die es zu analysieren gilt. Problematisch ist allerdings die Vielzahl möglicher Terme, so dass sehr schnell sehr viele m-elementige Termkombinationen entstehen können. Für jede dieser m-elementigen Kombinationen ist sicherzustellen, dass diese mindestens k mal im Datenbestand vorkommen.

2.4.2.4. Anonymisierung von standortbezogenen Diensten

Daten, die beim Einsatz standortbezogener Dienste (engl. Location-Based Services, LBS) anfallen, können personenbezogen sein. Aus diesem Grund fordert die EU-Direktive 2002/58/EC [Par02], dass der Nutzer eines solchen Dienstes einwilligen muss, bevor dessen exakte Position verarbeitet wird. Dieses Vorgehen birgt jedoch Probleme. Die Einwilligung wirkt pauschal, das heißt, einmal gegeben erlaubt sie die Verarbeitung der Standortdaten zu jedem (in der Einwilligung genannten) Zeitpunkt. Der Nutzer kann seine Einwilligung zwar widerrufen, dies ist im Allgemeinen jedoch nicht praktikabel. Möchte ein Nutzer kurzzeitig die Erhebung und Verarbeitung seiner Standortdaten unterbinden, zum Beispiel beim Betreten eines Krankenhauses, ist es wenig zielführend, dass der Nutzer seine Einwilligung widerruft und sich nach dem Verlassen des Krankenhauses neu anmeldet. Häufig ist ein derartiger Widerruf auch – entgegen dem Prinzip einen Medienbruch zu vermeiden – nur in Schriftform möglich. Außerdem erlaubt eine Einwilligung keine feingranulare Anpassung der Privatheitspräferenz des Nutzers an seine aktuelle Situation.

PETs für standortbezogene Dienste gibt es. Diese reichen von einfachen Schaltern zum Ein- / Ausschalten der GPS-Übermittlung [BD03], über Mechanismen, bei denen die Nutzer unterscheiden können, welche Gruppe von Personen Standortinformationen

einsehen darf. Das heißt, die Nutzer können definieren, wer die Standortdaten einsehen darf und sie können die Akkuratheit der Standortinformation beeinflussen [CSM+05], beispielsweise statt dem exakten Standortdatum einen Raum mehrerer möglicher Koordinaten preisgeben.

Während [BD03, CSM+05] das Einwirken des Nutzers erfordern, gibt es auch für standortbezogene Dienste den Ansatz der Anonymisierung. Das bedeutet, dass wann immer das Endgerät eines Nutzers dessen aktuelle Position an einen Anbieter übermittelt, die Anfrage ununterscheidbar von k anderen Anfragen gemacht wird.

Bei den Verfahren zur Anonymisierung unterscheiden wir zwei unterschiedliche Klassen: Die erste bezieht sich auf den Ort, an dem die Anonymisierung stattfindet. Der Ort kann entweder eine vertrauenswürdige Instanz sein, oder das Endgerät des Nutzers selbst. Beispielsweise sammelt in [GL04] ein 'protection broker' als vertrauenswürdige Instanz die Anfragen mehrerer Nutzer an einen Anbieter und anonymisiert diese, bevor er die anonymisierten Anfragen an den eigentlichen standortbezogenen Dienstgeber weiterleitet. In [KYS05] wird die Anonymisierung auf den Endgeräten selbst berechnet. Zweitens unterscheiden wir, ob es sich um einzelne, unabhängige Anfragen handelt oder um voneinander abhängige, kontinuierliche Anfragen. Letztere finden zum Beispiel dann statt, wenn das Endgerät immer neue Informationen zu dem aktuellen Umfeld anzeigen soll. Entsprechend sendet das Gerät kontinuierlich die Position des Nutzers an den Anbieter und erhält im Austausch vom Anbieter Informationen zu seinem Standort.

Unabhängige Anfragen Verfahren, die für einzelne, unabhängige Anfragen geeignet sind, umfassen unter anderem die Folgenden:

[KYS05] fügt in reale Anfragen an den Anbieter Anfragen für sogenannte 'Dummies' ein. Das sind imaginäre Nutzer, die imaginäre Anfragen an den Anbieter stellen. Der Anbieter soll auf diese Art nicht unterscheiden können, welches die Anfragen eines echten Nutzers sind und welche die imaginärer Nutzer. Der Anbieter antwortet auf jede dieser Anfragen. Die Antworten werden auf dem Gerät des Nutzers gefiltert, das die imaginären von der tatsächlichen Position zu unterscheiden weiß.

CliqueCloak [GL04] führt die Idee des räumlichen (engl. spatial) wie zeitlichen (engl. temporal) Verhüllens (engl. cloaking) ein. Dazu kann ein Nutzer pro Anfrage definieren, welche Anonymität (k) die Anfrage haben soll, zuzüglich einer räumlichen wie zeitlichen Toleranz. Diese Toleranzen erlauben auf der einen Seite die Anpassung des Anfrageraumes so, dass k Nutzer darin liegen. Außerdem ermöglichen sie die zeitliche Verzögerung der Anfrage, so dass die Wahrscheinlichkeit, dass ein anderer Nutzer ebenfalls eine Anfrage zu einem bestimmten Bereich absetzt, steigt. Anfragen, die das gleiche Informationsbedürfnis definieren und mindestens die gleiche Anforderung an die Anonymität haben, bilden die Cliquen.

Casper [MCA06] ergänzt die Idee der k-Anonymität um einen minimalen Bereich,

zu dem eine Anfrage gestellt werden darf.

Beispiel 8: Der Besuch eines Krankenhauses soll privat bleiben. Angenommen eine Anzahl Personen größer gleich k-1 befindet sich bereits in diesem Krankenhaus. Unter Verwendung von k-Anonymität würde eine Fläche an den Anbieter eines Dienstes übertragen, die die (nahezu exakte) Position des Krankenhauses widerspiegelt.

Ähnlich des beschriebenen Homogenitätsproblems wäre die Privatheit der Nutzer verletzt, trotz dass die Definition von k-Anonymität erfüllt ist. Der minimale Bereich schützt davor indem er vorgibt, wie klein ein angefragter Bereich minimal sein darf. Um die k Nutzer zu finden, die zusammengefasst werden sollen, unterteilt [MCA06] den betrachteten Raum in Zellen. Die Zellen bilden zusammen ein Gitter. Die Autoren definieren Zellen unterschiedlicher Größe, das heißt mehrere Gitter unterschiedlicher Granularität. Das grobgranularste Gitter besteht nur aus einer einzigen Zelle. Anschließend ordnen die Autoren die Gitter pyramidenartig an. Die Zelle ist der Bereich, der an den standortbezogenen Dienst übertragen wird. Um diese zu finden, ziehen die Autoren zuerst die feingranularste Zelle, in der der Nutzer liegt, heran und wandern anschließend die Pyramide der Gitter so lange zu der grobgranulareren Zelle nach oben, bis die Zelle neben der betrachteten Person k-1 weitere Personen umfasst.

Kontinuierliche Anfragen [KYS05, XC07, TM08] beschreiben Anonymisierungstechniken für kontinuierliche Anfragen und Trajektorien. Reine k-Anonymität reicht hier nicht aus, da die Anonymisierungen zum Zeitpunkt t und t+1 nicht unabhängig sind. In anderen Worten lässt das Wissen, ob ein Nutzer in einer bestimmten Region gewesen ist, Rückschlüsse darauf zu, wo er sich hinbewegt haben könnte.

Die Idee der Dummies [KYS05] findet auch für kontinuierliche Anfragen Anwendung. Schwierig ist es jedoch, solche imaginären Nutzer zu erzeugen, die auch zu einem realistischen Bewegungsprofil führen. Die Dummies müssen sich, damit ein Angreifer sie nicht gleich als solche erkennen kann, entlang von Straßen bewegen, sinnvolle Ziele ansteuern, sich mit einer sinnvollen Geschwindigkeit bewegen und vieles mehr. Identifiziert ein Angreifer einen Dummy, sinkt der Grad der Anonymität. Die Autoren stellen zwei Mechanismen vor, den für Bewegungen in der Nachbarschaft und den für Bewegungen in einer beschränkten Nachbarschaft. Letzterer berücksichtigt die Position anderer (echter) Nutzer bei der Berechnung von Dummies.

Die Ansätze von [GG03, XC07] basieren ebenfalls auf der Idee der 'Cloaking Areas', das heißt des Verhüllens von Anfragen. Insbesondere nutzt [XC07] das Konzept der Entropie und berücksichtigt die Wahrscheinlichkeit, mit der ein Nutzer innerhalb einer betrachteten Hülle liegt. Der Entropiewert einer Hülle kann dabei als die Menge korrelierenden Wissens interpretiert werden, die erforderlich ist, um einen Nutzer innerhalb der betrachteten Hülle zu identifizieren. Ist die Unordnung minimal, kann ein Angreifer einen Nutzer leicht identifizieren, ist sie maximal, kann er Nutzer innerhalb der Hülle nur mit gleicher Wahrscheinlichkeit identifizieren. Die Autoren stellen einen

Algorithmus vor, der performant die kleinste Hülle findet und gleichzeitig die Privatheit steigert – das heißt hier, ähnlich der Idee von [KYS05], die Verteilung der Nutzer berücksichtigt.

3. Analyse des Nutzer- und Anbieterverhaltens bei der Dienstnutzung

Das Web 1.0 und Web 2.0 bieten bequeme Möglichkeiten, Dienste von Behörden, Dienste zum Online-Einkaufen, zur Online-Unterhaltung oder dem sozialen Austausch zu nutzen. Die Anbieter dieser Dienste benötigen zur Diensterbringung personenbezogene Daten ihrer Nutzer – mit nicht absehbaren Folgen für die Privatheit des Einzelnen. Die gegenwärtige Datenschutzsituation stellt eine Herausforderung für die Privatheit nahezu jeden Bürgers der Gesellschaft dar.

In diesem Kapitel beschreiben wir eigene Vorarbeiten. Ausgewählte Erkenntnisse daraus untersuchen wir in Kapitel 4 und Kapitel 5 im Detail. Ziel ist es hier, in der Breite zu untersuchen, wie Nutzer mit Web 1.0- und Web 2.0-Diensten umgehen und umgekehrt, wie die Anbieter der Dienste dies mit den Nutzern tun. Diese Vorarbeit ist insbesondere durch das 2007 in Kraft getretene Telemediengesetz erforderlich geworden. Mit diesem Gesetz hat sich der Rechtsrahmen für Anbieter und Nutzer und somit auch der Ausgangspunkt für die Forschung im Kontext Privatheit und Datenschutz verändert.

Im Folgenden untersuchen wir den Status Quo nach der Gesetzesänderung. Wir konzentrieren uns dabei auf die beiden Schwerpunkte *Bewusstsein* (engl. awareness) der Nutzer und *Transparenz* der Anbieter. Wie wir in Beispiel 9 anhand Abbildung 3.1 exemplarisch zeigen werden, ist sowohl Bewusstsein als auch Transparenz erforderlich, möchte ein Nutzer die Kontrolle über seine personenbezogenen Daten behalten. Die Darstellung der Abbildung ist das Ergebnis der Begutachtung prominenter Dienste im Internet, die wir in Hinblick auf den Beitrag dieses Kapitels durchgeführt haben. Sie setzt sich zusammen aus drei Phasen der Dienstnutzung, mit abwechselnd erforderlichem Bewusstsein und erforderlicher Transparenz. Außerdem unterscheidet sie zwischen Nutzer-Nutzer- und Nutzer-Anbieter-Privatheit im Kontext des Web 2.0.

Beispiel 9: Man stelle sich einen Nutzer vor, der sich bei einem Web 2.0-Dienst anmelden möchte. In der *Registrierungsphase* braucht es aus Sicht der Privatheit zuerst die Transparenz des Anbieters, also welche Attribute der Anbieter bei der Registrierung wie, zu welchem Zweck und in welchem Umfang erhebt. Außerdem muss klar werden, welche Formen der Datenerhebung er einsetzt (Formulare, Cookies, . . .) und wie er die erhobenen Daten verarbeiten, nutzen oder weitergeben wird. Der Nutzer braucht ein ausgeprägtes Bewusstsein, um die Konsequenz

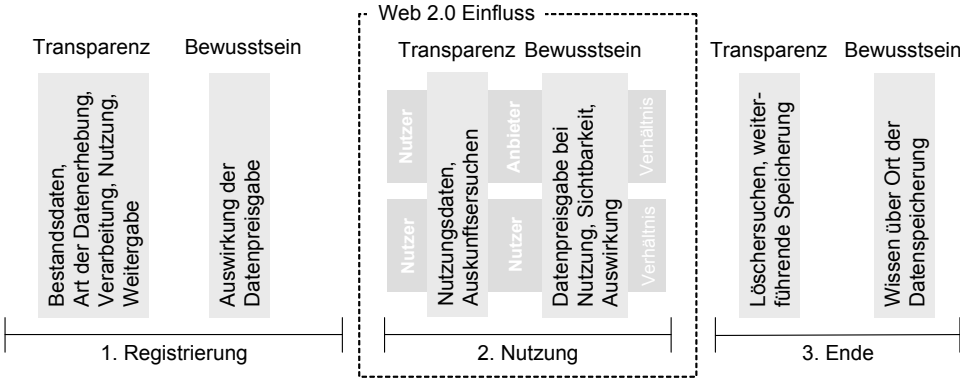


Abbildung 3.1.: Beispielhafte Nutzung eines (Web 2.0-) Dienstes

der Datenpreisgabe auf seine Privatheit abschätzen zu können. Für die Phase der *Dienstnutzung* muss der Anbieter wieder Transparenz schaffen, zum Beispiel angeben, welche Nutzungsdaten anfallen und wie er mit diesen verfährt, oder Auskunftsersuchen beantworten. Der Nutzer muss sich wiederum bewusst machen, welche Informationen er über die Daten der Registrierung hinaus preisgibt, beispielsweise in Foreneinträgen, und welche Auswirkungen die Preisgabe hat. Wie beschrieben gibt es neben der Nutzer-Anbieter Datenschutzpräferenz insbesondere mit dem Web 2.0 auch ein Nutzer-Nutzer Verhältnis. Der Nutzer muss sich der Sichtbarkeit der Daten für unterschiedliche Personenkreise und Unternehmen bewusst sein. Möchte der Nutzer zum *Ende* der Dienstnutzung seinen Zugang löschen, muss der Anbieter transparent machen, wie eine Löschung möglich ist, diese durchführen und auch auf Ausnahmen hinweisen, sollten die Daten beispielsweise zu Abrechnungszwecken nicht sofort gelöscht werden. Der Nutzer muss sich wiederum des Speicherorts bewusst sein, speziell nach einer Datenweitergabe durch den Anbieter.

Der Gesetzgeber, Diensteanbieter und die Forschung haben viele rechtliche sowie technische Mechanismen unter der Annahme vorgeschlagen und eingeführt, dass der Nutzer in der Lage ist, bewusst seine Privatheit zu schützen und er Bedrohungen für seine Privatheit wahrnimmt. Die Frage ist jedoch, ob diese Annahme aktuell überhaupt gültig ist. Ebenso ist es eine offene Frage, ob der Grad an Transparenz, so wie er aktuell bei Unternehmen vorherrscht, ausreicht, um zumindest dem bewussten Nutzer die Nachverfolgung seiner personenbezogenen Daten zu ermöglichen.

Wir werden im Folgenden Hypothesen motivieren, formulieren und evaluieren, die (i) das Bewusstsein der Nutzer und (ii) den Grad der Transparenz der Anbieter im Umgang mit personenbezogenen Daten betreffen. Zur Evaluierung nehmen wir dabei zwei Perspektiven ein, einmal mit Blick auf die Anbieter und einmal mit Blick auf die Nutzer. Hypothesen bezogen auf den Nutzer umfassen, dass Nutzer vergessen, wo sie welche Daten preisgegeben haben, Daten aus Versehen preisgeben und sich der Preisgabe

oftmals nicht bewusst sind. Bezogen auf den Anbieter stellen wir die Hypothese auf, dass der Fluss personenbezogener Daten oftmals intransparent ist, selbst wenn Gesetze diesbezüglich existieren. Zur Evaluation dieser Hypothesen wenden wir entsprechend der Perspektive eine unterschiedliche Methodik an. So führen wir eine Nutzerstudie mit technikgeschulten Teilnehmern durch, eine E-Mailumfrage unter Anbietern und untersuchen händisch Internetforen und Datenschutzerklärungen.

Unsere Ergebnisse zeigen [2], dass selbst technikgeschulte Personen nicht in der Lage sind, ihre Privatheit im Internet zu kontrollieren. Wir gehen aufgrund der Vorbildung unserer Teilnehmer der Studie davon aus, dass dieses Problem für weniger technologieaffine Personen und somit für weite Teile der Gesellschaft ebenso zutrifft. Die Harmonisierung der Gesetzgebung innerhalb der EU erlaubt den Schluss, dass unsere Ergebnisse auch für andere EU-Länder gelten.

3.1. Motivation der Hypothesen

In diesem Abschnitt beschreiben wir eine Auswahl von Beobachtungen, die wir bei unserer Recherche im Internet gemacht haben. Sie stellen zugleich die Motivation für unsere Hypothesen bezüglich des (nicht vorhandenen) Nutzerbewusstseins und der (In-) Transparenz der Datenverarbeitung der Anbieter dar.

3.1.1. Beobachtungen bezüglich Nutzer (Nutzerperspektive)

Dieser Abschnitt beschreibt die Beobachtungen mit Blick auf den Nutzer.

Beobachtung 1 *Nutzer wechseln Online-Dienste häufig und vergessen die Registrierungen früherer Zugänge.*

Viele Internetnutzer registrieren sich bei Diensten, löschen jedoch die nicht mehr gebrauchten Zugänge nicht. Solche Registrierungen erfolgen beispielsweise bei Downloadplattformen auf der Suche nach Softwarewerkzeugen (Freeware, Shareware, etc.) oder beim Ausprobieren neuer Dienste. Die Hürde, solche Dienste auszuprobieren, ist (wie gewünscht) niedrig. Die einmal preisgegebenen personenbezogenen Daten verbleiben zumeist jedoch ohne Löschung für immer beim Diensteanbieter.

Beobachtung 2 *Viele Nutzer sind zufrieden mit der Existenz einer Datenschutzerklärung, lesen diese aber nicht sorgfältig.*

Nationales sowie internationales Recht zwingen einen Anbieter, die Nutzer über seine Datenschutzpraktik in einer Datenschutzerklärung zu unterrichten. Diese Datenschutzerklärungen sind häufig umfangreich, schwierig zu lesen und verschleiern Sachverhalte

in unspezifischen Formulierungen.

Beobachtung 3 *Nutzer vergessen häufig, dass Freundeslisten, Chat-Protokolle etc. im Internet für lange Zeit öffentlich sichtbar sind.*

Die Gesellschaft, insbesondere die junge Generation, nutzt das Internet zum Pflegen sozialer Kontakte (engl. socializing) und ersetzt dadurch oftmals sogar den klassischen gegenseitigen Austausch in Form eines physischen Treffens der Beteiligten [BM04].

Beobachtung 4 *Viele Nutzer sind inkonsequent beim Einsatz von Pseudonymen, das heißt, trotz des Einsatzes eines Pseudonyms geben sie Daten preis, die ihre Identifikation erlauben.*

Die EU [Par95] (Artikel 6) fordert die Anbieter (implizit) auf, so wenig personenbezogene Daten wie möglich zu sammeln. Viele Anbieter, zum Beispiel Forenbetreiber, bieten aus diesem Grund die Möglichkeit der pseudonymisierten Anmeldung und / oder Dienstnutzung. Trotz der Möglichkeit einer pseudonymisierten Dienstnutzung geben viele Nutzer während der Nutzung doch wieder solche Daten preis, die eine Identifikation erlauben. Beispiele sind die Preisgabe einer Telefonnummer in einem Forum, um sich mit Freunden verabreden zu können, oder die Angabe einer öffentlich sichtbaren Versandadresse.

3.1.2. Beobachtungen bezüglich Anbieter (Anbieterperspektive)

Die folgenden Beobachtungen beziehen sich auf den Anbieter.

Beobachtung 5 *Unternehmen haben ein ökonomisches Interesse, große Mengen personenbezogener Daten über lange Zeit zu speichern.*

Das Geschäftsmodell vieler Dienste umfasst das Erheben und Speichern detaillierter Informationen über die Nutzer. Ziel ist zum Beispiel die Personalisierung des Dienstes oder die Personalisierung von Werbung (siehe auch Kapitel 2.1). Besonders kritisch wird die Konzentration dieser Daten im Falle von Unternehmensübernahmen oder wenn Unternehmen Daten anderen Unternehmen der gleichen Unternehmensgruppe zugänglich machen.

Beobachtung 6 *Die Verknüpfung personenbezogener Daten ermöglicht es Anbietern, umfassende Profile der Nutzer zu erstellen.*

In der Datenschutzforschung ist die Verknüpfung personenbezogener Daten ein bekanntes Problem. Beispielsweise können Registrierungsdaten mit Daten aus öffentlichen Quellen kombiniert werden [Gol06, Spi07] oder mit Datenbanken anderer Unternehmen der Unternehmensgruppe. Neue Herausforderungen entstehen dabei durch

Dienste wie Spock.com oder Yasni.de. Diese sogenannten Personensuchmaschinen verknüpfen alle öffentlich zugänglichen Quellen zu einer Person im Internet zu einem umfassenden Profil.

3.2. Nutzer- und Anbieterstudie

In diesem Abschnitt stellen wir unsere Hypothesen auf und validieren diese. Wir werden zeigen, dass Nutzer nicht in der Lage sind, die Möglichkeiten der existierenden Gesetzgebung zu nutzen. Wir führen für die Auswertung vorab zwei Definitionen bezüglich der Struktur der preisgegebenen Daten ein. Abschnitt 3.2.1 beschreibt die zur Validierung der Hypothesen eingesetzte Methodik. Abschnitt 3.2.2 motiviert jede Hypothese und stellt unser Ergebnis vor.

Personen geben im Internet Daten zu ihrer Identität strukturiert (Beobachtung 1, 2, 4) und unstrukturiert (Beobachtung 3, 5, 6) preis.

Definition 2 *Strukturierte Preisgabe personenbezogener Daten bedeutet, dass ein Nutzer die sensiblen Daten (Attributwerte) explizit einem Attribut zuordnet.*

Die strukturierte Preisgabe personenbezogener Informationen in (Anmelde-) Formularen stellt unmittelbar den Bezug zwischen dem preisgegebenen Attributwert und der Semantik des Wertes, dem Attributtyp, her. Ein Beispiel ist die Unterscheidung zwischen Vor- und Nachnamen, insbesondere bei fremdländischen Namen. Bei der strukturierten Preisgabe gibt der Nutzer durch die Wahl des entsprechenden Eingabefeldes diese Information explizit mit an. Dies vereinfacht es dem Anbieter zum Beispiel, personenbezogene Daten mit Daten aus anderen Quellen zu verknüpfen.

Definition 3 *Unstrukturierte Informationspreisgabe bezieht sich auf Situationen, in denen ein sensibles Datum nur implizit mit einem Attributnamen korreliert werden kann.*

Beispiele für die unstrukturierte Preisgabe personenbezogener Daten umfassen Einträge in Blogs, Produktrezensionen und Gästebucheinträge; kurz, jede Preisgabe, bei der Nutzer Klartext in natürlicher Sprache hinterlassen. Ohne den Einsatz von Methoden zur automatischen Verarbeitung natürlicher Sprache (natural language processing techniques, NLP), die den Text in strukturierte Daten umwandeln, kann ein Anbieter diese Daten nicht verarbeiten. Außerdem führen die erwähnten Techniken häufig zu fragwürdigen und unpräzisen Ergebnissen.

3.2.1. Methodik

Es ist unser Ziel, die Interaktionen von Nutzern und Anbietern einmal mit Blick auf den Nutzer und einmal mit Blick auf den Anbieter zu untersuchen. Dazu benötigen wir eine unterschiedliche Methodik zur Validierung unserer Hypothesen. Zur Adressierung der Nutzerperspektive führen wir eine Nutzerstudie unter unseren Kollegen und Studenten durch. Weiter untersuchen wir das Nutzerverhalten in öffentlichen Internetforen. Wir untersuchen die Anbieterperspektive mit Hilfe einer E-Mailumfrage, der händischen Untersuchung der Datenschutzerklärungen und einer Analyse der Registrierungsprozesse der Anbieter.

Nutzerstudie Um Ergebnisse zu erzielen, die für eine große Gruppe von Personen repräsentativ sind, haben wir uns entschieden, unsere Studie mit erfahrenen Internetnutzern durchzuführen. Wenn wir bestätigen, dass selbst diese Personengruppe bei der Nutzung von Diensten im Internet ihre Privatheit nicht sicherstellen kann, haben wir einen starkes Indiz, dass dies auch für Nutzer mit geringerer Erfahrung im Umgang mit Online-Diensten gilt. Wir haben uns für Informatikstudenten entschieden. Diese sind vertraut im Umgang mit Online-Diensten und nutzen das Internet als Ressource sowohl für die Organisation ihrer Ausbildung als auch des sozialen Lebens.

Vor der Hauptstudie haben wir die Methodik mit 18 Mitarbeitern der Arbeitsgruppe getestet und die Umfrage entsprechend der gewonnenen Erfahrungen angepasst. Für die endgültige Studie haben wir in Summe 95 Freiwillige aus unserer Vorlesung eingeladen. Wir haben keine Auszahlung angeboten. 18 Studenten haben eine Teilnahme aus Zeit- oder anderen Gründen abgelehnt und 5 nachdem sie den Inhalt der Studie erfahren hatten. Somit sind 72 Teilnehmer (17% weiblich, 83% männlich) verblieben. Die Studie wurde in unserem Labor durchgeführt. Von Anfang an war das Thema Datenschutz bekannt. Wir gehen somit davon aus, dass wir eher optimistische Werte erhalten, was den Umgang mit personenbezogenen Daten angeht. Oder anders ausgedrückt, können die Personen, obwohl sie das Ziel der Studie kennen, nicht bewusst mit ihren Daten umgehen, so gilt dies sicherlich auch dann, wenn ihr Fokus nicht auf Datenschutz liegt. Das Ausfüllen des Fragebogens dauert ca. 20 Minuten. Die Teilnehmer wurden zum Ausfüllen des Fragebogens separiert. Jedem Teilnehmer stand bei einer unklaren Frage jederzeit ein Ansprechpartner zur Verfügung.

Untersuchung des Nutzerverhaltens in Internetforen Die unstrukturierte Preisgabe personenbezogener Daten bezieht sich unter anderem auf öffentliche Foren, die die Mitglieder unter Verwendung eines Pseudonyms nutzen. Wir haben die internationalen Foren der New York Times und der Washington Post händisch analysiert. In beiden Foren haben wir uns auf die aktivsten Fäden (Threads) des jeweiligen Forums konzentriert. Um eine breite Menge von Nutzern zu erhalten, sind die Themen der Fäden breit aufgestellt, zum Beispiel Elektrotechnik, Blogger, Gesundheit oder Essen. Für die New

York Times haben wir 440 Beiträge analysiert, für die Washington Post 200 Beiträge von 54 Nutzerprofilen.

Untersuchung von Anbietern Um mehr über die Datenschutzpraktiken von Unternehmen zu erfahren, haben wir die Datenschutzerklärungen und Registrierungsprozesse von vielfrequentierten Online-Anbietern untersucht. Da NLP Techniken auf Datenschutzerklärungen bisher nicht erfolgreich angewendet werden können (vergleiche Kapitel 2.4.1), untersuchen wir auch hier die Anbieter händisch. Wir haben dazu auf der einen Seite die Top-25 Anbieter untersucht, die Teilnehmer in unserer Nutzerstudie genannt haben – wir haben zu diesem Zweck eine entsprechende Frage im Fragebogen gestellt – und auf der anderen Seite die Top-10 Anbieter nach der Rangliste von ComScore[1]. In Fällen, in denen ComScore nur einen Konzern aber nicht die einzelnen Mitglieder listet, haben wir das größte Unternehmen des Konzerns als Repräsentanten ausgewählt.

E-Mailumfrage Gesetzliche Regelungen zwingen einen Anbieter, seine (potentiellen) Nutzer über die Erhebung, Verarbeitung und Nutzung personenbezogener Daten zu informieren. Dies geschieht in vielen Fällen jedoch nur sehr vage und unter der Verwendung unspezifischer Formulierungen. Kann der Nutzer aufgrund der Formulierungen den Umgang mit seinen Daten nicht erschließen, so liegt in der Regel ein Verstoß vor. Wir haben die Unternehmen mit unspezifischen Formulierungen in ihren Datenschutzerklärungen über E-Mail kontaktiert. Wir haben die E-Mails personalisiert, das heißt, wir haben die jeweils unklaren Textpassagen aus den Datenschutzerklärungen extrahiert und in der entsprechenden E-Mail ausgewiesen. Den Kontext einer Studie haben wir den Unternehmen nicht mitgeteilt. Die Anbieter konnten in einem Zeitfenster von 2 Monate antworten.

3.2.2. Hypothesen und Evaluierung

Unsere Beobachtungen in Abschnitt 3.1 haben Herausforderungen für den Datenschutz im Internet aufgezeigt. In diesem Abschnitt werden wir diese Herausforderungen methodisch anhand von fünf Hypothesen untersuchen. Dabei behalten wir die Unterscheidung in die Perspektive des Nutzers und die des Anbieters bei.

3.2.2.1. Datenschutzbewusstsein der Nutzer

These 1 Bewusstsein-1 *Nutzer können die strukturierte Preisgabe von Informationen nicht nachvollziehen.*

[1] `http://de.sys-con.com/read/467066.htm` April 2010

Motivation: Beobachtung 1 (Nutzer wechseln Online-Dienste häufig und vergessen die Registrierungen früherer Zugänge) und Beobachtung 5 (Unternehmen haben ein ökonomisches Interesse, große Mengen personenbezogener Daten über lange Zeit zu speichern) motivieren unsere erste Hypothese. Wir wollen insbesondere wissen, ob sich die Nutzer früherer Registrierungen bewusst sind. Wenn unsere Hypothese stimmt, wenden Nutzer keinen Aufwand für die Einhaltung ihrer Privatheit auf.

Evaluation: Wir können unsere Hypothese belegen, wenn wir zeigen, dass sich Teilnehmer an eine deutlich kleinere Menge von Anbietern erinnern als an die Menge, bei der sie tatsächlich registriert sind. Zu diesem Zweck haben wir die folgenden Fragen mit in unseren Fragebogen aufgenommen:

Frage 1 *Bei welchen Anbietern haben Sie in der Vergangenheit personenbezogene Daten hinterlassen? Listen Sie ALLE Anbieter mit Name / URL explizit auf.*

Die Studienteilnehmer haben die Frage 1 zweimal beantworten müssen. Zuerst haben wir sie aufgefordert, alle Anbieter niederzuschreiben, bei denen sie personenbezogene Daten preisgegeben haben. Am Ende des Fragebogens haben wir die gleiche Frage erneut gestellt, haben jedoch Anbieterkategorien als Beispiele mit einem jeweiligen Repräsentanten angegeben, zum Beispiel 'Soziale Netzwerkseite, StudiVZ', 'Instant Messenger, MSN/Skype' etc. Offensichtlich haben wir keine vollständige Liste aller möglichen Anbieter vorlegen können. Ist die Differenz der Anzahl Registrierungen, an die sich die Nutzer von sich aus erinnern und derer, an die sie sich erst mit unserer Hilfe erinnern können, trotzdem schon deutlich, so bestärkt das die Prüfung unserer Hypothese jedoch nur.

Die Umfrage hat gezeigt, dass Nutzer bei einer Vielzahl von Anbietern registriert sind. Von 70 Teilnehmern (zwei wollten diese Frage nicht beantworten) haben wir 259 Namen von Anbietern erhalten, bei denen die Teilnehmer registriert sind.

Abbildung 3.2 zeigt die kumulative Verteilungsfunktion für die Anzahl der Registrierungen mit und ohne Unterstützung durch die Angabe der Anbieterkategorien. Die horizontale Achse gibt die mittlere Anzahl der Registrierungen an. Die vertikale Achse zeigt die Anzahl der Teilnehmer. So kann man beispielsweise ablesen, dass sich 60% unserer Teilnehmer an weniger als 8 Registrierungen erinnern, wenn sie nicht durch die Beispielkategorien und Beispielrepräsentanten pro Kategorie unterstützt werden. Der Unterschied zwischen den beiden Kurven zeigt, dass die Anzahl der Registrierungen, die den Nutzern bewusst sind, weit hinter der tatsächlichen Anzahl von Registrierungen zurückliegt.

Die maximale Anzahl von Registrierungen ist 29, die maximale Anzahl erinnerter Registrierungen ohne Unterstützung 21. Ohne Unterstützung haben sich 45 Teilnehmer (64%) an nur weniger als 50% ihrer Registrierungen erinnert. Zusammengefasst zeigt unsere Evaluierung, dass sich viele Internetnutzer der Preisgabe ihrer personen-

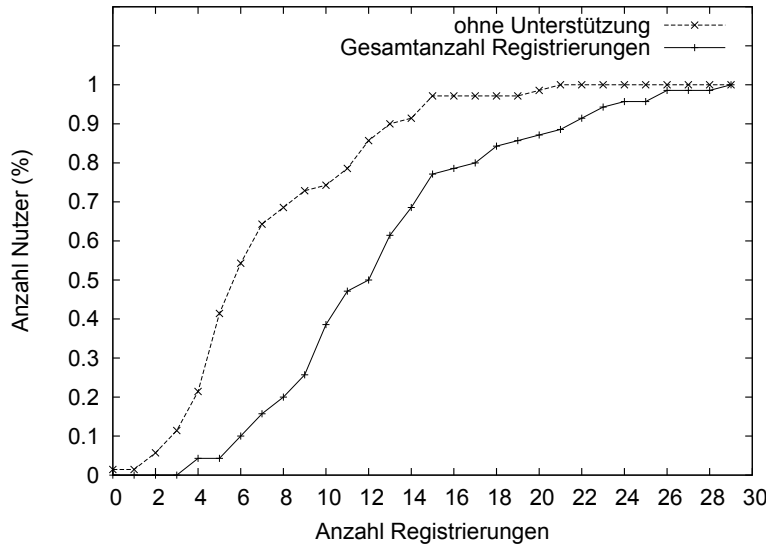


Abbildung 3.2.: Kumulative Verteilungsfunktion der Anzahl Registrierungen / Nutzer

bezogenen Daten bei einer Vielzahl von Anbietern nicht bewusst sind.

Am Ende des Fragebogens haben wir alle Teilnehmer gefragt:

Frage 2 *Waren Sie sich der Tatsache, sich an so viele Preisgaben von personenbezogenen Daten nicht mehr erinnern zu können, bewusst?*

Im Ergebnis haben 51 (73%) ausgesagt, dass sie solch eine große Differenz erwartet haben. 17 (24%) sind überrascht gewesen, zwei haben die Frage nicht beantwortet.

Die Differenz zwischen den realisierten und der erst nach Unterstützung erinnerten Registrierungen sowie das Desinteresse der Teilnehmer, über die Preisgabe ihrer Daten Protokoll zu führen, unterstützt unsere Hypothese.

These 2 Bewusstsein-2 *Nutzer geben versehentlich personenbezogene Daten in unstrukturierter Form preis.*

Motivation: Personen nutzen das Internet zum gegenseitigen Austausch (engl. socializing) zwischen Freunden und für andere Formen der privaten Kommunikation. Das bedeutet in vielen Fällen, dass zumindest Fragmente dieses Austausches im Internet, zum Beispiel in Foren oder in sozialen Netzwerkseiten, sichtbar sind (Beobachtung 3). Außerdem impliziert Beobachtung 4, dass Internetnutzer manchmal ihre wahre Identität preisgeben. Nutzer tun dies selbst dann, wenn sie die Wahl haben, ein Pseudonym einzusetzen. Diese Hypothese spiegelt wieder, dass viele Nutzer sich der Möglichkeit,

dass andere Personen umfangreiche Profile über sie erstellen könnten, nicht bewusst sind.

Evaluierung: Unsere Hypothese hält, wenn wir Foren finden, die pseudonym genutzt werden können, bei denen jedoch Nutzer unnötigerweise personenbezogene Daten preisgeben. Zu diesem Zweck haben wir zwei Foren, erstens von der New York Times (NYT) und zweitens von der Washington Post (WP) untersucht. Die NYT fordert von ihren Nutzern eine Registrierung, erlaubt es den Nutzern aber, zu jedem Beitrag ein neues Pseudonym einzusetzen. Wenn die Nutzer diesen Mechanismus korrekt einsetzen, können andere Forenbesucher weder die betroffene Person identifizieren, noch können alle Einträge eines Nutzers diesem durch eine Verknüpfung über das Pseudonym zugeordnet werden. Die WP erfordert ebenfalls eine Registrierung. Sie ordnet jeden Beitrag einem Nutzerprofil zu. Jeder Nutzer des WP Forums kann die Profile aller anderen Nutzer einsehen, inklusive dem Geschlecht, Alter, Heimatort, Lieblingszitat und Arbeitsbeschreibung.

Die Untersuchung von 440 Forenbeiträgen der NYT hat zu 206 Beiträgen geführt (47%), die die Autoren mit ihrem Vornamen unterschrieben haben. Bei 120 (27%) Einträgen ist sowohl der Vor- als auch der Nachname angegeben. Einige davon (16) haben sogar ihren Heimatort beziehungsweise Staat oder ihre persönliche Webseite mit angegeben, zum Beispiel ‘John Public, FL’. Die Untersuchung bei der WP hat zu anderen Ergebnissen geführt. Nur drei der 54 untersuchten Nutzer haben zusätzlichen Informationen neben ihrem Pseudonym in ihrem Profil hinterlegt.

Obwohl wir nicht ausschließen können, dass ein paar der NYT Beiträge unter erfundenen Identitäten getätigt worden sind, sehen wir unsere Hypothese durch die schiere Anzahl unnötiger Preisgaben personenbezogener Informationen bestätigt. Wie jedoch die Untersuchung bei WP gezeigt hat, hängt der korrekte Einsatz, das heißt das Bewusstsein eines Nutzers bezüglich eines PETs, von weiter zu untersuchenden Faktoren ab. Wie ersichtlich ist, wurde der Mechanismus bei der NYT, der eigentlich einen besseren Schutz der Privatheit ermöglichen sollte, falsch eingesetzt, wohingegen der wenig ausgefeilte Mechanismus bei der WP, insbesondere die Existenz von Autorenprofilen, zu besseren Ergebnissen geführt hat. Die untersuchte soziale Gruppe, die Nutzerschnittstelle und die im Forum behandelten Themen beeinflussen die Bereitschaft von Personen, personenbezogene Daten preiszugeben.

These 3 Bewusstsein-3 *Nutzer sind sich identifizierender Attribute oder Attributkombinationen nicht bewusst.*

Motivation: Entsprechend der Beobachtung 5 und 6 haben viele Unternehmen ein wirtschaftliches Interesse an der Sammlung und Verknüpfung personenbezogener Daten. Das Zusammenführen unterschiedlicher Attribute zu Attributkombinationen, die ein Individuum eindeutig beschreiben, gefährden dessen Privatheit. Beispielsweise ist bekannt, dass die Verknüpfung der Postleitzahl, des Geschlechts und des Geburtstages die

Tabelle 3.1.: Anzahl Identifikatoren pro Teilnehmer

Anz. Schlüssel	Anz. Teilnehmer	Anz. Teilnehmer (%)
0	17	24%
1	22	31%
2	17	24%
3	8	11%
4	3	4%
5	4	6%
6	1	1%

Identifikation von 63% der US Bevölkerung erlaubt [Gol06] (siehe auch Kapitel 1.2). Beispiele von einzelnen Attributen, die in Korrelation mit Hintergrundwissen identifizierend sein können, sind die Ausweisnummer oder die Matrikelnummer. Diese Hypothese spiegelt wieder, dass sich viele Personen der Preisgabe identifizierender Attribute beziehungsweise Attributkombinationen nicht bewusst sind. Besonders wichtig ist dieser Sachverhalt bei der Kombination von personenbezogenen Daten aus unterschiedlichen Informationsquellen zu einem umfassenden Profil.

Evaluation: Wir können unsere Hypothese validieren, indem wir zeigen, dass viele Internetnutzer sich der Tatsache nicht bewusst sind, dass bestimmte Attribute oder Attributkombinationen die eindeutige Identifikation einer Person zulassen. Zu diesem Zwecke haben wir die folgende Frage gestellt:

Frage 3 *Bitte nennen Sie alle Attribute oder Attributkombinationen, die es einem Anbieter ermöglichen könnten, Sie zu identifizieren.*

Da die Gruppe der Teilnehmer Informatikstudenten unserer Vorlesung gewesen sind, können wir davon ausgehen, dass sie mit der Idee von Identifikatoren vertraut sind. Sie kennen diese zum Beispiel als Schlüssel in Datenbanken oder Verknüpfungen zwischen unterschiedlichen Relationen etc. Obwohl der Fragebogen, um das Ergebnis nicht zu beeinflussen, keine Beispielidentifikatoren enthalten hat, erwarten wir, dass zumindest offensichtliche Identifikatoren wie Vorname / Nachname, Telefonnummer etc. genannt werden.

Tabelle 3.1 und Tabelle 3.2 zeigen die Ergebnisse der Evaluierung dieser Hypothese. Tabelle 3.1 zeigt die Anzahl der unterschiedlichen Attribute und Attributkombinationen, die die Teilnehmer als Identifikatoren klassifiziert haben. Wir haben die Teilnehmer aufgefordert, so viele Attributkombinationen wie möglich zu nennen. Tabelle 3.2 weist die von den Teilnehmern häufiger als dreimal genannten Attributkombinationen explizit aus.

Überraschenderweise waren 17 Personen (24%) nicht in der Lage, auch nur einen

Tabelle 3.2.: Identifizierende Attributkombinationen

Attributkombination	Anzahl
Nachname, Vorname	16
Nachname, Vorname, Geburtstag	14
Nachname, Vorname, Adresse	10
Bankleitzahl, Kontonummer	8
E-Mail	5
Telefonnummer	4
Kreditkartennummer	4
Adresse	4
Personalausweisnummer	4
Pseudonym, Geburtstag	4
Nachname, Adresse	4
IP-Adresse	3
Matrikelnummer	3

Identifikator zu nennen. 22 Personen (31%) haben exakt einen Identifikator genannt, 17 (24%) zwei. 16 Teilnehmer haben angegeben, dass die Kombination aus Vor- und Nachname in ihrem Fall für die Identifikation ausreichend sei. 14 Teilnehmer haben die Kombination aus Vor- und Nachname um das Geburtsdatum erweitert. Es hat uns überrascht, dass vier Personen ihr Pseudonym und Geburtstag als Identifikator angegeben haben. Die Teilnehmer erklärten aber, dass viele Systeme automatisch das Geburtsjahr einem Pseudonym nachstellen würden, wenn das eigentliche Pseudonym bereits von einem anderen Nutzer verwendet wird. Außerdem würden sie das Pseudonym in so vielen Kontexten benutzen, dass es als Identifikator ausreichend sei. Weiter haben ebenfalls nur je vier Personen die Kreditkartennummer, die Telefonnummer, die Adresse, die Personalausweisnummer und die Kombination aus Nachname und Adresse als Identifikator genannt.

Wir können nicht ausschließen, dass einzelne Teilnehmer die Frage nicht verstanden haben. Selbst dann zeigen die Ergebnisse jedoch, dass sich sehr viele Personen über identifizierende Attribute und die Verknüpfbarkeit von Daten keine Gedanken machen. Wir sehen unsere Hypothese somit als bestätigt.

3.2.2.2. Transparenz der Anbieter

These 4 Transparenz *In vielen Fällen wissen Nutzer nicht oder können gar nicht herausfinden, welche Daten bei der Dienstnutzung erhoben werden.*

Motivation: Beobachtung 5 hat auf das wirtschaftliche Interesse von Unternehmen hin-

gewiesen, personenbezogen Daten zu sammeln. Der Gesetzgeber wirkt dem insofern entgegen, als er ohne Einwilligung des Nutzers die Erhebung nur solcher Daten zulässt, die zur Diensterbringung offensichtlich erforderlich sind. Konkret fordert er ihn zur Datensparsamkeit auf. Die folgende Hypothese bezieht sich auf das Problem, dass viele Nutzer zur Zeit der Registrierung noch gar nicht vorhersehen können, ob sich der Anbieter an das Gebot der Datensparsamkeit hält.

Evaluation: Wir können unsere Hypothese validieren, wenn wir eine markante Anzahl relevanter Anbieter finden, die ihre Nutzer nicht vorab über die preiszugebenden Daten informieren. Um dies zu untersuchen, haben wir die Anbieter händisch analysiert.

Wir haben dabei drei unterschiedliche Klassen von Registrierungsprozessen identifiziert. Die einfachste Variante (1) ist die Registrierung mittels genau einer Seite beziehungsweise eines Formulars (engl. single page registration, SPR). Der Anbieter fragt mittels eines Formulars nur hier und alle Daten auf einmal ab. Die zweite Form (2) ist die Registrierung und Erhebung personenbezogener Daten auf mehreren Formularen oder Seiten (engl. multi-page registration, MPR). Bei MPR muss der Nutzer beispielsweise zum Lesen von Foreneinträgen eine E-Mailadresse und ein Pseudonym hinterlegen. Möchte er Beiträge schreiben, erscheint ein neues Formular, über das sein realer Name erhoben wird. Der Radiosender SWR3 erhebt zum Beispiel (Stand 7. April 2010) eine Handynummer beim Schreiben in Foren, um – wir haben bei dem verantwortlichen Datenschutzbeauftragten nachgefragt – einen Identifikator des Autors im Falle von rassistischen Kommentaren zu besitzen. Eine hybride Variante sind (3) Registrierungsprozesse, bei denen der Anbieter einen Nutzer durch mehrere Dialoge führt (engl. wizard drive registration, WDR). Hier wird alle Information auf einmal erhoben, jedoch auf mehreren hintereinandergeschalteten Seiten.

Während es SPR den Nutzern erlaubt, alle erforderlichen Informationen auf einen Blick zu erfassen, so verstecken MPR und WDR Attribute auf Folgeseiten. Gemäß des Transparenzgebots muss ein Anbieter den Nutzer über die Daten, die er erhebt, informieren. Dies gilt insbesondere dann, wenn der Anbieter sie zu unterschiedlichen Zeitpunkten erhebt.

Abbildung 3.3 stellt die Registrierungsprozesse dar, die wir identifiziert haben. 10 der 20 Anbieter sammeln alle erforderliche Information auf einer Seite (SPR). 5 Anbieter nutzen MPR, und die gleiche Anzahl Anbieter WDR. Die Abbildung zeigt weiter, dass nur zwei der Anbieter vorab angeben, zum Beispiel in der Datenschutzerklärung, welche weiteren Daten ein Nutzer angeben muss. Weitere personenbezogene Daten fallen beim Bezahlen an, bei einer Bestellung etc. Nur drei der WDR Anbieter weisen aus, dass weitere Seiten folgen, auf denen der Nutzer personenbezogene Daten angeben muss. Somit kommen selbst einige der hier untersuchten großen Anbieter ihren Informationspflichten nicht nach und verletzen das Gesetz. Dies bestätigt zugleich unsere

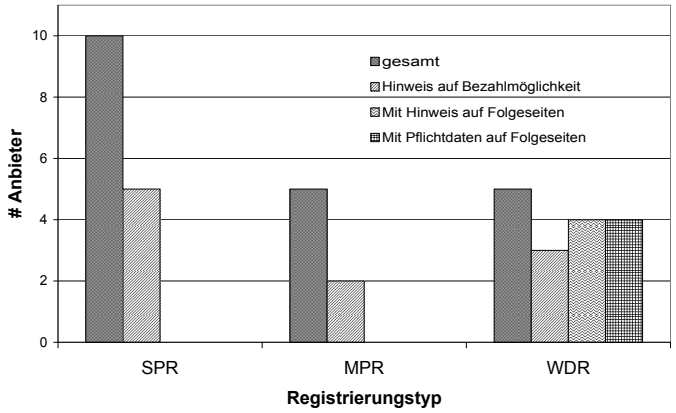


Abbildung 3.3.: Einsatz unterschiedlicher Registrierungstypen

Hypothese.

These 5 Transparenz-2 *Nutzer können den Fluss von Informationen zwischen Unternehmen nicht nachvollziehen.*

Motivation: Große Unternehmen, die wie Google viele unterschiedliche Dienste anbieten oder die wie die Holtzbrink Gruppe viele Unternehmenstöchter haben, geben in ihren jeweiligen Datenschutzerklärungen häufig an, personenbezogene Daten an abhängige oder verbundene Unternehmen weiterzugeben. Gemäß Beobachtung 6 hat dies einen klaren Einfluss auf die Privatheit der Nutzer: Nutzer geben unter Umständen unterschiedliche Informationen bei verschiedenen Töchtern einer Unternehmensgruppe an, die dann von dem Unternehmen zu umfassenden Profilen kombiniert werden können. Außerdem ist es ausgesprochen schwierig, alle Unternehmen und Beteiligungen einer Unternehmensgruppe zu identifizieren.

Evaluation: Um unsere Hypothese zu validieren, haben wir direkten Kontakt mit den Anbietern aufgenommen. Anbieter, die angeben keine Daten weiterzugeben, haben wir ausgeschlossen. Gleiches gilt für Unternehmen, die angeben, vor jeder Weitergabe der Daten bei der betroffenen Person eine Einwilligung einzuholen.

Tabelle 3.3 zeigt die Anbieter mit unspezifischen Formulierungen bezüglich der Datenweitergabe durch das Unternehmen. Unsere E-Mailumfrage untersucht drei Arten von unspezifischen Formulierungen: Wir haben die Anbieter (A) nach ihrer Unternehmensstruktur gefragt und (B) wer die in der Datenschutzerklärung genannten 'verwand-

Tabelle 3.3.: Weitergabe personenbezogener Daten

Anbieter	Kategorie	(A/B/C)	Addr.	Antwortzeit, unbeantw.
Amazon	Shop	A/B	nein	keine Antwort
Otto	Versandhändler	A/B	ja	3 Tage, B unbeantw.
eBay	Auktionsplattform	A/B/C	nein	13 Tage
GMX	Freemailer	A/B/C	ja	7 Tage
Web.de	Freemailer	A/C	nein	keine Antwort
MSN	Freemailer	A/B	ja	1 Tag, A/B unbeantw.
Google	Freemailer	A/B	ja	keine Antwort
Skype	Messenger	A/C	ja	2 Tage
AOL.de	Comunities	A/B	ja	keine Antwort
Yahoo.de	Comunities	A	ja	16 Tage, A unbeantw.
Sat1	Forum	B	ja	7 Tage

ten Unternehmen' sind, an die Daten weitergegeben werden. Einige Anbieter geben weiter an, unter 'besonderen Umständen' Daten weiterzugeben. Diese vage Formulierung (C) wollten wir uns auch von den Anbietern konkretisieren lassen. Für den Betroffenen ist ein vollständiges Verständnis von A, B und C erforderlich, um die Auswirkungen einer Registrierung auf die Privatheit abschätzen zu können.

Alle Anbieter in Tabelle 3.3 geben an, Daten weiterzugeben. Alle geben Daten innerhalb der Unternehmensgruppe weiter, ohne explizit die Empfänger zu nennen (A). Sat1 gibt die Daten ebenfalls weiter, nennt aber deren Empfänger. Skype und Web.de geben an, Daten zu keinem Unternehmen außerhalb der Unternehmensgruppe weiterzugeben. Amazon, Otto, eBay, GMX, MSN, Google und AOL sagen, dass sie Daten zu verbundenen Unternehmen weitergeben, zum Beispiel zu Dienstleistern, um den geforderten Dienst erbringen zu können (B). eBay, GMX, Web.de und Skype geben an, Daten unter speziellen Umständen weiterzuleiten, ohne diese zu nennen. Wir haben alle Anbieter mit der folgenden E-Mail angeschrieben:

> Sehr geehrte Damen und Herren,
>
> ich kann die Auswirkungen der folgenden drei Formulierungen in Ihrer Datenschutzerklärung nicht abschätzen. Erstens kann ich nicht erkennen, welche Unternehmen (an die Sie Daten weitergeben) Teil Ihrer Unternehmensgruppe sind. Zweitens verstehe ich nicht, was verwandte / verbundene Unternehmen sind. Ich würde Sie bitten, mir zwei Listen, einmal Ihrer Unternehmensgruppe und einmal mit den verwandten/verbundenen Unternehmen, zukommen zu lassen. Drittens geben Sie an, Daten unter 'besonderen Umständen' oder bei 'berechtigtem Interesse' weiterzugeben. Ich

bitte Sie, diese Formulierung zu konkretisieren.

Mit freundlichen Grüßen
(Unterschrift)

Innerhalb der nächsten 16 Tage haben wir von 7 Unternehmen Antwort erhalten. Web.de hat uns eine automatische Antwortmail geschickt mit einer anzurufenden Telefonnummer. Bei unserem Anruf wurde uns mitgeteilt, dass die Daten an niemanden weitergegeben worden seien. Wir haben darauf hingewiesen, dass ihre Datenschutzerklärung etwas Anderes aussagt. Die Mitarbeiter von Web.de konnten ihre Datenschutzerklärung jedoch selbst gar nicht finden und baten uns, diese ihnen per E-Mail zukommen zu lassen. Wir sind dieser Bitte nachgekommen, haben aber leider keine Antwort mehr erhalten.

Weitergabe innerhalb der Unternehmensgruppe (A) MSN und Yahoo! haben uns geantwortet, dass sie keine Liste der Unternehmensgruppe herausgeben dürften. Offensichtlich unzureichend hat Web.de geantwortet. eBay hat uns aufgefordert, bei all ihren Marktplätzen das jeweilige Impressum zu lesen, um die Unternehmen zu identifizieren, die verantwortlich sind – bei der wachsenden Anzahl von Marktplätzen eine fast unmögliche Aufgabe.

Nur Otto, GMX und Skype sind unserer Aufforderung gefolgt und haben die Mitglieder der Unternehmensgruppe ausgewiesen.

Weitergabe an verbundenen Unternehmen (B) eBay gibt an, die Dienstleister aufgrund von Firmengeheimnissen nicht nennen zu können. GMX und Sat1 haben eine detaillierte Liste der Partner geliefert. Otto hat die Kundennummer angefordert, um die verbundenen Unternehmen zu nennen, an die Daten weitergegeben wurden. Wie haben die Kundennummer übermittelt, daraufhin jedoch keine Antwort erhalten.

Weiterleitung unter besonderen Umständen eBay hat in seiner Antwort erklärt, dass die Bezeichnung ‘legitimes Interesse’ bei Bedarf die Weiterleitung zu allen eBay-Marktplätzen bedeute, zum Beispiel, wenn ein Marktplatzteilnehmer den Kontakt zu einem Kunden verloren hat. Skype hat geantwortet, dass ‘spezielles Interesse’ die ‘Interessen von Skype oder einem anderen Unternehmen der Gruppe’ bedeutet. Unternehmen, die nicht in der Gruppe sind, hat Skype ausgeschlossen. Web.de hat nicht geantwortet.

Während der Auswertung haben wir zwei weitere Erkenntnisse erlangt, die unsere Hypothese unterstützen.

Erkenntnis 1 *Viele Anbieter weisen keinen Kontakt für Datenschutzfragen aus.*

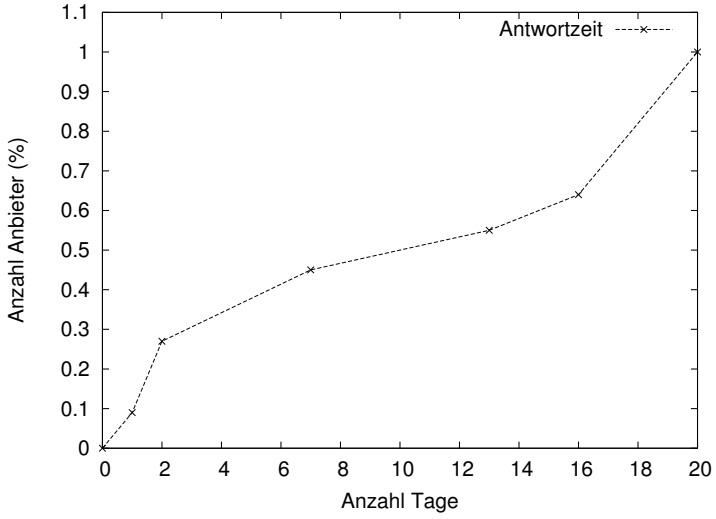


Abbildung 3.4.: Antwortzeit der Anbieter

Um den Datenschutzbeauftragten der Unternehmen zu kontaktieren, haben wir die Datenschutzerklärung, das Impressum etc. durchsucht. MSN hat ein Kontaktformular angeboten, bei dem man in die Datenschutzerklärung einwilligen muss, obwohl man unter Umständen gerade zum Thema Datenschutz eine Frage stellen möchte. Drei Anbieter (Amazon, eBay und Web.de) haben keine Kontaktinformation zu Datenschutzfragen ausgewiesen.

Erkenntnis 2 *Viele Anbieter reagieren langsam auf Datenschutzfragen.*

Abbildung 3.4 zeigt die Anzahl der Werktage, die die Anbieter zur Beantwortung unserer Anfragen gebraucht haben. Amazon, Google und AOL haben gar nicht geantwortet. Wie anhand der Abbildung ersichtlich ist, haben 27% der Anbieter innerhalb einer Woche geantwortet.

Zusammengefasst bedeuten unsere Ergebnisse, dass es für die betroffene Person schwierig bis unmöglich ist herauszufinden, welche Unternehmen nach einer Registrierung Kenntnisse über die preisgegebenen personenbezogenen Daten haben. Dies lässt sich in zwei Probleme zusammenfassen: Erstens kommen die Unternehmen ihrer Pflicht, den Nutzer zu informieren, nicht korrekt nach. Zweitens machen es viele Anbieter ihren Kunden schwer, sie zu kontaktieren. Sie reagieren nur langsam auf Anfragen im Vergleich zu der Zeit, in der sich Teilnehmer registrieren können, in der ein Anbieter ein Produkt liefern kann etc. Dies lässt den Schluss zu, dass es für Auskunfts-

ersuchen keine etablierten Unternehmensprozesse gibt. Zuletzt sind die Antworten auf Anfragen vielfach unzureichend.

3.3. Zusammenfassung und Diskussion

In der bis hierhin vorgestellten Vorarbeit haben wir zentrale Herausforderungen für zukünftige PETs im Internet identifiziert und exemplarisch untersucht. Diese umfasssen sowohl Herausforderungen zum Schutz des Datenaustausches zwischen Nutzern, als auch zwischen dem Nutzer und dem Anbieter. Unsere Ergebnisse zeigen, dass (1) Nutzer sich der unkontrollierten Preisgabe personenbezogener Daten nicht bewusst sind. So spiegeln unsere Hypothesen Bewusstsein 1 – 3 wider, dass viele Personen die Preisgabe ihrer Daten nicht kontrollieren. Sie protokollieren die Preisgabe nicht und geben Daten versehentlich preis. Weiter zeigen wir aber auch (2), dass selbst der bewusste Nutzer mangels Transparenz nicht in der Lage ist, den Fluss seiner Daten zu verfolgen. Die Hypothesen Transparenz 1 und 2 sagen uns, dass viele Unternehmen nicht angeben, welche Daten bei der Registrierung erforderlich sind. Außerdem befolgen sie die Informationspflichten auf eine Art und Weise, dass sie der Intention (Telos) des Gesetzgebers nicht genügen. Ein Beispiel ist die Weitergabepraxis innerhalb einer Unternehmensgruppe. Die daraus resultierende Ungewissheit für den Nutzer hat auch Auswirkungen auf andere Datenschutzgesetze: Wie soll zum Beispiel ein Nutzer seine Daten löschen können, wenn er nicht nachvollziehen kann, wer alles diese Daten erhalten hat? Außerdem scheint es weder für Auskunfts- noch Löschersuchen etablierte Unternehmensprozesse zu geben. Die Konsequenz für fast alle Gesellschaftsschichten ist hoch: Viele Gesetze und Mechanismen zum Schutz von Privatheit gehen von einem bewusst handelnden Nutzer aus. Diese ersten Ergebnisse zeigen jedoch, dass das unrealistisch ist. Außerdem kann der Gesetzgeber aktuell keinen transparenten Fluss von personenbezogene Daten sicherstellen. Der Einsatz von zukünftigen Data-Mining-Techniken, beispielsweise von Personensuchmaschinen, wird diesen Effekt noch verstärken. Nach unserem Kenntnisstand berücksichtigen existierende Datenschutzmechanismen nicht alle hier aufgezeigten Problematiken.

Resultierende Anforderungen Aus dieser Vorarbeit lassen sich erste, allgemeine Anforderungen an PETs ableiten. Diese sind relevant für die strukturierte als auch die unstrukturierte Preisgabe von Daten sowie für den Web 1.0- und Web 2.0-Kontext:

A1 Ein PET sollte die preisgegebenen Daten und den Ort der Preisgabe lokal auf dem Rechner des Nutzers protokollieren und dem Nutzer jederzeitigen Zugriff darauf geben.

A2 Das PET sollte im Hintergrund arbeiten und den Aufwand des Nutzers auf ein Minimum reduzieren.

A3 Das PET sollte den Nutzer warnen, wenn er Daten preisgibt, die mit anderen Daten korreliert werden können.

A4 Das PET sollte ohne Einfluss des datenerhebenden Unternehmens funktionieren.

Unsere Ergebnisse rufen nach der Entwicklung neuer vertrauenssteigernder Techniken und der Durchsetzung existierender Gesetze. Es ist unser Forschungsziel, (i) obige Anforderungen an PETs systematisch zu entwickeln, anhand konkreter Implementierungen und mit menschlichen Probanden. Außerdem (ii) ist es unser Ziel, durch die Entwicklung eines PETs die Durchsetzung existierender Gesetze zu verbessern.

4. PETs im Web 2.0 zum Schutz zwischen Nutzern

Wie die Ergebnisse unserer Vorarbeit (Kapitel 3) gezeigt haben, gibt es zwei zentrale Überlegungen bei dem Entwurf eines Mechanismus zum Schutz der Privatheit (PET): *Erstens*, soll das PET den Schutz zwischen Nutzern untereinander oder zwischen einem Nutzer und einem Anbieter fördern? *Zweitens*, soll das PET bei der Durchsetzung von Transparenz auf Anbieterseite oder bei der Förderung des Bewusstseins der Nutzer ansetzen?

In diesem Kapitel widmen wir uns dem Schutz zwischen Nutzern. Es ist unser Ziel zu identifizieren, welchen Anforderungen ein PET genügen muss, um den Privatheitspräferenzen von Web 2.0-Nutzern zu entsprechen. Wir unterscheiden die Anforderungen dabei in (i) inhaltliche Anforderungen, (ii) Anforderungen an die Benutzbarkeit und (iii) funktionale Anforderungen. (i) beschreibt, 'was', 'wann', 'vor wem' geschützt werden soll, also den Inhalt der Präferenz. (ii) gibt an, wie ein PET geartet sein muss, damit es nicht nur theoretisch funktioniert, sondern auch vom Nutzer korrekt eingesetzt werden kann. (iii) bezieht sich auf die Umsetzung der Anforderungen aus (i) und (ii), also das 'wie'. Es umfasst dabei auch generelle Anforderungen wie A1-A4 aus Abschnitt 3.3, also zum Beispiel, dass ein PET ohne Einfluss des Diensteanbieters funktionieren soll.

Um die Anforderungen zu identifizieren, untersuchen wir im Folgenden real funktionierende Dienste anhand von Nutzerstudien. Die Untersuchung zweier unterschiedlicher Dienste ist erforderlich, da der Begriff des Web 2.0 eine Vielzahl von Dienstvarianten umfasst, die Personen in verschiedenen Kontexten einsetzen.

Die beiden untersuchten Dienste sind kollaborative Suchmaschinen (Abschnitt 4.1) und standortbezogene Dienste (Abschnitt 4.2). Diese sind repräsentativ für zwei große Klassen von Web 2.0-Diensten, die kurz davor sind, den Alltag zu durchdringen.

4.1. PETs und Kollaboration bei Suchmaschinen

4.1.1. Privatheitsprobleme kollaborativer Suchmaschinen

Kollaborative Suchmaschinen (Collaborative Search Engines, CSE) erweitern die Suche im Internet durch die Möglichkeit, dass Nutzer Suchterme, Suchergebnisse und angeklickte Links untereinander austauschen. Beispiele kollaborativer Suchmaschinen sind I-SPY [SBB+05], MUSE und MUST [RJK08], SearchTogether [MH07] und Fireball LiveSearch[1], wenn auch letzteres nur sehr eingeschränkte Funktionalität hat. CSEs lassen Personen, die zum Beispiel regelmäßig Wissen unterschiedlichster Quellen konsultieren müssen, den Aufwand beim Suchen untereinander synchronisieren. Nutzer, die erfahren im Umgang mit Suchmaschinen sind, können unerfahrene Nutzer mit Hilfe von CSEs unterstützen und anleiteten, und die Internetnutzer können im Stil einer Web 2.0-Anwendung kollaborativ Recherchen betreiben.

Beispiel 10: Die Planung eines Familienausfluges könnte wie folgt aussehen: Jeder sucht sich einen Ort seiner Wahl, Übernachtungsmöglichkeiten, Ausflugsziele etc. Jeder Vorschlag wird jedem anderen Familienmitglied mitgeteilt. Der Empfänger setzt dann wiederum auf diesen bisherigen Suchergebnissen auf, um sich selbst ein Bild des Ziels machen zu können und teilt Änderungswünsche oder Alternativen wiederum jedem mit.

Bisher wurde eine Suche wie in Beispiel 10 aber auch beim wissenschaftlichen Arbeiten, beim Einkaufen und der Suche nach Produkten händisch durchgeführt [MH07]. So tauschen die Nutzer diese Ergebnisse beispielsweise per E-Mail untereinander aus. CSEs unterstützen den Nutzer beim Austausch der Informationen bis hin zu einem automatisierten Informationsaustausch. Da Suchanfragen und die angeklickten Ergebnisse jedoch die Gewohnheiten, Interessen, sozialen Kontakte sowie die Intention des Nutzer verraten können [Dat08], sind CSEs aus Sicht der Privatheit problematisch.

Konkret gibt es zwei Probleme: Ein strukturbedingtes Problem und ein inhaltsbedingtes. Strukturbedingtes Problem bedeutet, dass durch die langfristige Speicherung der Suchanfragen und der zentralen Verarbeitung der Daten in der CSE unter Umständen neue Privatheitsprobleme entstehen. [4] schlägt einen Ansatz vor, der diesem Problem entgegenwirkt. Das inhaltsbedingte Problem bezieht sich, wie der Name sagt, auf die unter den Nutzern ausgetauschten Inhalte. Wir konzentrieren uns hier auf diesen zweiten Aspekt.

Aktuell existierende CSEs geben entweder an, dass jede Suchinformation für andere Nutzer sichtbar ist, und / oder überlassen es dem Nutzer, manuell Leute in sogenannte Such-Sitzungen (engl. Sessions) einzuladen. Jede Session umfasst einen bestimmten Ausschnitt von Informationen, von denen der Eingeladene profitiert [MH07]. Der Idee von CSEs liegt jedoch auch ein automatisierter Austausch von Daten zugrunde, zum Beispiel im Sinne von Anwendungen wie friendfeed.com, bei der Nutzer relevante In-

[1] `http://fireball.de`

halte abonnieren können. Wäre bei jedem Austausch von Inhalten, wie Suchanfragen zwischen Nutzern, ein manuelles Einwirken des Nutzers erforderlich, würden CSEs nur schlecht funktionieren. Die Akzeptanz von CSEs wird daher von der Angemessenheit integrierter Mechanismen zum Schutz der Privatheit (Privacy-Enhancing Technologies, PETs) abhängen.

Es ist schwierig zu bestimmen, was ein angemessenes und somit auch effektives PET ausmacht: Das Ziel von CSEs ist es, den gegenseitigen Austausch zwischen Nutzern zu fördern. Da existierende Gesetze jedoch hauptsächlich den Umgang vom Staat und von Unternehmen mit personenbezogenen Daten beschreiben, bieten Gesetze nur unzureichende Lösungsmöglichkeiten. Auch technische Lösungen wie P3P [CLM] sind nur für Anbieter-Nutzer-Verhältnisse konzipiert. CSEs erfordern jedoch Wissen über die Privatheitspräferenzen zum Schutz der Privatheit der Nutzer untereinander, als auch den Schutz vor Dritten. Beide, andere Nutzer wie Anbieter, könnten versuchen, Wissen aus den Suchanfragen und angeklickten Links der CSE-Nutzer abzuleiten. Es ist davon auszugehen, dass bekannte Privatheitspräferenzen von Nutzern gegenüber Anbietern davon verschieden sind. Zusätzlich haben vorangegangene Studien gezeigt, dass Privatheit ein sehr von Emotionen getriebenes Thema ist und Nutzer in Laborexperimenten und / oder Fragebögen nicht notwendigerweise ihre wahren Privatheitspräferenzen preisgeben [SGB01, CKMD06].

Wir wollen die Privatheitspräferenzen von Nutzern beim Einsatz kollaborativer Suchmaschinen herausfinden und Erkenntnisse erlangen, wie ein PET geartet sein muss, damit Nutzer ihre Präferenzen vollständig spezifizieren können. Konkret lauten die wesentlichen Fragen, die wir mit dieser Studie beantworten möchten [3, 4]:

F1 Was sind die *Parameter* einer Strategie, die die Präferenzen des Nutzers widerspiegeln?
F2 Welche *Personengruppen*, zum Beispiel Freunde, Verwandte oder Bekannte, adressieren die Teilnehmer mit ihrer Strategie?
F3 Gibt es eine allgemeine *Struktur*, auf die sich die meisten Strategien abbilden lassen, das heißt eine leicht automatisiert auswertbare Darstellung?
F4 Unterscheiden Nutzer in ihren Privatheitspräferenzen zwischen den *Suchanfragen* und den angeklickten *Suchergebnissen*?

Zur Beantwortung dieser Fragen haben wir (1) eine kollaborative Suchmaschine entworfen und implementiert. Unsere CSE verfügt über eine Schnittstelle, über die Nutzer ihre Präferenzen im Umgang mit ihren Daten definieren können. Wir bezeichnen diese im Folgenden als *Strategien*. Basierend auf unserer Implementierung haben wir (2) eine Nutzerstudie durchgeführt. Anhand von 27 Studenten der Informatik im Hauptstudium haben wir unsere Forschungsfragen evaluiert. Da es sich bei CSEs um eine neue Art der Anwendung handelt, haben wir unsere Teilnehmer intensiv mit dem Konzept und der Nutzung der CSE vertraut gemacht. Die Teilnehmer haben ihre Präferenzen in

natürlicher Sprache formuliert. Natürliche Sprache schließt aus, dass die Teilnehmer aufgrund technischer Limitierungen oder mangelnder Ausdrucksmächtigkeit, von zum Beispiel einer formalen Sprache, eingeschränkt sind [SSP09].

Unsere Evaluierung gibt Entwicklern von PETs interessante Einblicke in die Privatheitspräferenzen von Nutzern: Nutzer sind besonders besorgt darüber, was Freunde, Kollegen und die Familie über ihre Suchanfragen herausfinden können. Das ist besonders interessant im Kontext der Ergebnisse von [MH07], da das genau die Gruppen sind, mit denen CSE Nutzer kollaborieren möchten. Wir haben außerdem herausgefunden, dass die Privatheitspräferenzen von Nutzern sich auf eine einfache Struktur abbilden lassen, sich jedoch auf unterschiedliche Arten von Bedingungen und Abhängigkeiten beziehen. Darüber hinaus definieren Nutzer auch reziproke Präferenzen. Das heißt, dass sie die Auswirkung ihrer Strategie von dem Verhalten des Empfängers der Inhalte abhängig machen. Dies können ähnliche Interessen sein, ob der Teilnehmer kurz oder lang angemeldet ist oder ob der Empfänger ähnliche Daten ebenfalls preisgeben würde. Dieses Ergebnis ist erwähnenswert, da wir uns keiner Spezifikationssprache für Privatheitspräferenzen bewusst sind, die diesen Aspekt berücksichtigt. Zuletzt haben wir beobachtet, dass die Teilnehmer in ihren Strategien nur selten zwischen Suchanfragen und angeklickten Links aus dem Suchergebnis unterscheiden. Sie definieren jedoch Abhängigkeiten zwischen Suchanfragen und Links, so dass Anfragen nur preisgegeben werden dürfen, wenn Links zu bestimmten URLs angeklickt werden.

Wir werden in Abschnitt 4.1.2 unsere Methodik und zentralen Entwurfsentscheidungen der Studie darstellen. In Abschnitt 4.1.3 präsentieren wir unsere Studienergebnisse, zusammen mit einer einfachen Sprache zur Repräsentation der identifizierten Privatheitspräferenzen. Abschnitt 4.1.4 fasst die Ergebnisse zusammen.

4.1.2. Methodik der Nutzerstudie

Im Folgenden beschreiben wir unsere wesentlichen Entwurfsentscheidungen (Abschnitt 4.1.2.1) sowie Aufbau und Durchführung der Studie (Abschnitt 4.1.2.2).

4.1.2.1. Entwurfsentscheidungen

Dem Aufbau unserer Studie liegen fünf zentrale Entwurfsentscheidungen zugrunde:

Kompetente Teilnehmer Um den Informationsfluss von Anfragen und angeklickten Suchergebnissen innerhalb einer CSE verstehen zu können, bedarf es umfangreichen Wissens aus dem Kontext Informationssysteme. Unsere Studienteilnehmer sind Studenten im Hauptstudium mit einem Schwerpunkt auf Informationssystemen. Somit sind sie auch erfahrene Nutzer im Umgang mit dem Internet und dem Suchen und Finden von Informationen, also eine Zielgruppe von CSEs.

Schulung im Umgang mit der CSE Aktuell existieren CSEs nur als Forschungsprototyp oder Implementierungen in einem frühen Stadium. Wir können also nicht davon ausgehen, dass unsere Teilnehmer tiefgreifende Kenntnisse im Umgang mit CSEs haben. Um aussagekräftige Ergebnisse zu erzielen, müssen die Teilnehmer umfassendes Wissen über die verwendete Technologie haben [Bab07]. Wir haben selbst eine CSE implementiert und die Implementierungsdetails, den Quellcode, das Datenbankschema und die Testdaten offengelegt. Zum besseren Verständnis des Informationsflusses, haben die Teilnehmer Teile der CSE nachentwickelt. Um das Ergebnis nicht zu verfälschen, haben wir die Teilnehmer jedoch nicht zu Privatheitsbedrohungen und dem PET-Einsatz geschult.

Klartext Strategien Die automatisierte Auswertung einer Strategie erfordert ihre Formulierung in einer strukturierten, gegebenenfalls formalen Art und Weise. Das Problem ist, dass dadurch Einschränkungen genau der Strategien entstehen, deren Gestalt wir eigentlich mit unserer Studie erst erheben möchten, um anschließend eine strukturierte Darstellung zu schaffen. In sozialen Netzwerkseiten haben die Nutzer beispielsweise eine nur sehr beschränkte Flexibilität beim Spezifizieren der Privatheitspräferenzen und können auch nur zwischen einer beschränkten Menge von Strategien wählen [SSP09]. [PMTW08] stellt eine Sprache vor, jedoch ohne zu wissen, was die Nutzer ausdrücken möchten. Außerdem sind formale Darstellungen für viele Nutzer nicht intuitiv. Unter Umständen drücken die Nutzer nur aus was sie ausdrücken können, nicht aber, was sie wirklich ausdrücken möchten [Ack00]. Wir haben uns aus diesem Grund entschieden, die Strategien in Klartext zu erheben.

Beobachtung des Nutzerverhaltens Wir wissen von [Bab07], dass das Verhalten von Nutzern (i) abhängig von der verwendeten Technologie ist und (ii) Personen in Fragebögen ein anderes Verhalten an den Tag legen als beim realen Einsatz der Technologien [SGB01]. Aus diesem Grund haben wir unsere reale Anwendung mit einer Schnittstelle zur Definition der Privatheitspräferenzen versehen. Während die kollaborative Suchmaschine voll funktionsfähig ist, können die Privatheitspräferenzen in natürlicher Sprache nicht automatisiert verarbeitet werden. Die Teilnehmer haben die Strategien beim Einsatz der Suchmaschine immer dann formuliert, wenn ihnen die Preisgabe der aktuellen Suche widerstrebt hat. Das heißt auch, dass Nutzer jeweils mehrere Strategien definieren können, zwischen denen sie abhängig von ihrer aktuellen Situation wechseln. Für schützenswerte Suchen haben sie die Kollaboration deaktivieren können.

Zusammenfassen von Nutzern in Gruppen Soziale Netzwerke im Web 2.0 sind in der Regel sehr groß. Beispielsweise haben die Nutzer von MySpace im Mittel 115 Freunde und somit mehrere tausend Freundesfreunde [SSP09]. Nutzer unserer CSE definieren aus diesem Grund ihre Strategien gegenüber Personengruppen, wobei eine Gruppe auch aus nur einer Person bestehen darf, wie zum Beispiel 'Ehefrau'.

Tabelle 4.1.: Freischaltung der CSE-Komponenten nach Experimentphasen

Komponenten nach Phasen	P1	P2	P3	P4
Standardsuche	✓	✓	✓	✓
Ähnliche Anfragen		✓	✓	✓
Durchsuchen von Anfragen			✓	✓
Skype Schnittstelle			✓	✓
Strategie-Schnittstelle				✓

Wir beschreiben nun, wie wir diese Entwurfsentscheidungen umgesetzt haben. Informationen zu der Realisierung der CSE-Anwendung befinden sich in Anhang A.

4.1.2.2. Aufbau und Durchführung

Unsere Studie besteht aus fünf Phasen, von denen jede drei Wochen gedauert hat. In Phase P1-P3 haben wir die Teilnehmer im Umgang mit der CSE sowie den erhobenen und über die CSE ausgetauschten Daten geschult.

Die Teilnehmer haben Aufgaben bearbeitet, die den Umgang mit der verwendeten Datenbanktechnologie, zum Beispiel DDL, SQL, PL/SQL und Trigger zeigen. Nach der Bearbeitung dieser Aufgaben haben die Teilnehmer alle Datenbankrelationen, Beziehungen zwischen diesen, sowie Operationen, die die CSE einsetzt, mindestens einmal selbst erstellt und somit auch verstehen können. In der Phase P4 haben wir die Schnittstelle zum Erstellen der Strategien freigeschaltet. Anschließend haben die Teilnehmer ihre Strategien über diese Schnittstelle definiert. In Phase P5 haben wir abschließende Kontrollfragen gestellt, um Seiteneffekte, wie zum Beispiel eine falsche Verwendung oder ein falsches Verständnis der CSE, auszuschließen (siehe dazu auch Abschnitt 4.2.4.3). Damit jeder Nutzer ein umfassendes Verständnis der CSE erlangt, haben wir die einzelnen Komponenten eine nach der anderen freigeschaltet. Tabelle 4.1 gibt einen Überblick der freigeschalteten Komponenten für jede Experimentphase.

Jeder der Schritte führt immer weitere kollaborative Funktionalität ein. Das heißt, die Teilnehmer haben jede der Funktionen getrennt wahrgenommen sowie die damit einhergehende Auswirkungen auf die Datenverarbeitung und die Speicherung in der Datenbank. Für die Phasen P1-P3 haben die Teilnehmer Punkte erhalten, die wir am Ende zu einer Note zusammengeführt haben und die die Teilnehmer zum Bestehen der Veranstaltung benötigt haben. Dies ist ein klarer Anreiz für die Teilnehmer gewesen, sich beim Training in den ersten Phasen intensiv zu engagieren. Für Phase P4, die Definition der Strategien, haben wir, um das Ergebnis nicht zu beeinflussen, keine Bepunktung vorgenommen. Um die Teilnehmer in der abschließenden Umfrage zu motivieren (P5), haben wir zwei Amazon-Gutscheine unter den Teilnehmern verlost.

Einführungsphase (P1) Das Ziel der ersten Phase ist es gewesen, ein grundlegendes Verständnis der Funktion der CSE zu vermitteln. Wir haben zu diesem Zweck eine Präsentation vorbereitet, die die Funktionalität der CSE beschreibt und jeden Teilnehmer individuell nach vorgegebenen Themen aus dem Bereich 'Sport' suchen lässt. Bis hier ist der einzige Unterschied zu der Suche mit Google die vorherige Registrierung an unserer CSE gewesen. Layout und Funktion von Google sind unverändert geblieben.

Anschließen haben die Teilnehmer die Aufgabe bekommen, kleine Programme zu schreiben, die die Suchdaten, das heißt die Daten, die durch die Suchen der Teilnehmer selbst entstanden sind, auf unterschiedliche Art verwenden (SQL, PL/SQL etc.). Die Teilnehmer haben dabei Suchanfragen anderer Teilnehmer identifizieren müssen, die ähnlich zu ihren eigenen Anfragen sind, Statistiken über die Anfragen erstellt und Nutzer mit ähnlichen Suchprofilen ermittelt. Auch hier haben sie die Technologien eingesetzt, die von der CSE selbst verwendet werden. Außerdem haben die Teilnehmer Komponenten der CSE nachentwickelt.

Anfragephase (P2) Anschließend haben wir die Teilnehmer in die ersten kollaborativen Aspekte der CSE eingeführt. Wir haben in die Google-Oberfläche ein Fenster eingeblendet, dass ähnliche Anfragen und angeklickte Links anderer Nutzer der CSE anzeigt – an dieser Stelle jedoch zunächst, ohne dass der Name des Nutzers mit den ähnlichen Anfragen sichtbar ist. Dieses Verhalten ist ziemlich genau das, welches I-SPY [SBB+05] anbietet.

Die Teilnehmer haben die Anfragen anderer Nutzer und deren angeklickte Links genutzt, um schneller die ihnen gestellten Aufgaben beantworten zu können. Beispielsweise haben sie so untereinander interessante Seiten mit Code-Beispielen ausgetauscht, Hilfe-Dokumente referenziert etc. Wir haben darüber hinaus die Protokollierung aktiviert, über die wir erkennen können, welcher Teilnehmer auf welche Suchanfrage oder Link eines anderen Nutzers geklickt hat. Die Teilnehmer konnten die von unserer CSE protokollierten Informationen in der Datenbank und deren Verarbeitung einsehen.

Nutzer Bewusstsein (P3) In dieser Phase haben wir die Identitäten der Nutzer explizit gemacht, das heißt, die Namen der Teilnehmer öffentlich sichtbar neben den eingeblendeten ähnlichen Suchanfragen und Links angezeigt.

Die Teilnehmer haben ein Instant-Messenger-Plugin für Skype entwickelt. Über das Plugin haben sie die Anfragen anderer Nutzer abonnieren können. Sucht ein Nutzer mit Hilfe der CSE, werden dessen Suchanfragen automatisiert an alle Skype-Abonnenten verschickt. Jede dieser Nachrichten beinhaltet die Anfrage, angeklickte Links, den Namen des Absenders und die Zeit der Anfrage. Auf diesem Weg können die Teilnehmer mittels Skype kollaborieren und gegenseitig Informationen austauschen.

Strategiedefinitionphase (P4) In dieser Phase haben wir die Teilnehmer aufgefordert, ihre Strategien zum Schutz ihrer Privatheit über die Schnittstelle der CSE zu formulieren. Die Teilnehmer haben in Klartext definiert, in welchem *Kontext* sie mit *wem welche Anfragen und Links* austauschen würden.

Da die Strategien selbst ebenfalls private Informationen beinhalten können [YW03, ZL08], haben wir die Strategiedefinitionen den anderen Teilnehmern nicht zugänglich gemacht. Um die Teilnehmer nicht zu überfordern und um eine strukturierte Evaluierung zu ermöglichen, haben wir die Teilnehmer die Strategien in drei Schritten definieren lassen: (1) Strategien zum Schutz von Anfragen, (2) Strategien zum Schutz von Links und (3) die Möglichkeit, nach drei Wochen die Strategien zu überarbeiten. Für jede Strategie haben wir (i) nach der Situation gefragt, wann diese angewendet werden soll, (ii) nach der Nutzergruppe, auf die sich die Strategie bezieht, und (iii) ob anderen Teilnehmern durch die Strategie der Zugang zu den Daten ermöglicht oder verboten werden soll. Dies ist allerdings nur als Richtlinie gedacht. Durch die Verwendung von Klartextstrategien haben die Teilnehmer jede beliebige Strategie definieren können, die sie zum Ausdrücken ihrer Privatheitspräferenzen benötigen.

Abschlussumfrage (P5) Wir haben die Studie mit einer Abschlussumfrage beendet. Wir haben die Teilnehmer dabei, um die Qualität unserer Ergebnisse sicherzustellen, (i) Kontrollfragen zum Einsatz der CSE, Aufbau etc. gefragt. Außerdem (ii) haben wir zur besseren Interpretation unserer Ergebnisse ihre generelle Einstellung zum Thema Privatheit abgefragt, um zum Beispiel Datenschutzfundamentalisten [ACR99] auszuschließen.

4.1.3. Evaluation

Im Folgenden beschreiben wir die Ergebnisse unserer Studie [3, 4], das heißt, wir analysieren die Strategien, die die Teilnehmer definiert haben.

Insgesamt haben wir 247 Strategiedefinitionen von 27 Nutzern erhoben. Davon haben 142 Strategien den Schutz von Suchanfragen und 105 Strategien angeklickte Links betroffen. Zuerst analysieren wir die Strategien bezüglich der Anfragen. Wir untersuchen, (1) wie wir die Strategien strukturieren können, (2) den Kontext, auf den sie sich beziehen, (3) welche sozialen Gruppen benannt werden und (4) die Art der verwendeten Prädikate. Unter einem Prädikat verstehen wir zum Beispiel '[_] darf Inhalt sehen', wobei [_] anschließend zum Beispiel durch die Personengruppe 'Freunde' ersetzt wird. Anschließend betrachten wir, wie sich die Strategien für die angeklickten Links unterscheiden.

4.1.3.1. Struktur der Strategie

Wir haben herausgefunden, dass wir die in Klartext formulierten Strategien auf die folgende allgemeine Struktur abbilden können:

[ALWAYS | IF <*conditions*>] [DO NOT] DISCLOSE <*objects*> [TO <*groups*>]

Bedingungen (conditions), Objekte (objects) und Gruppen (groups) sind aus einem oder mehreren Termen zusammengesetzt. Verbunden sind sie mit einem logischen UND beziehungsweise ODER. Eine Beispielstrategie wäre 'Wenn ich am Arbeiten bin ODER in der Zeit zwischen 7:00 und 18:00, gib meiner Familie UND meinen Freunden KEINE Suchanfragen und Links preis'. Einige Strategien sind diesem Aufbau von Anfang an gefolgt, die anderen haben wir in diese Form überführen können.

Manche Strategien haben die Preisgabe der Daten pauschal zugelassen oder abgelehnt ('ALWAYS DO...'). Die Mehrheit der Strategien hat sich jedoch auf einen der folgenden Bedingungen bezogen:

1. *Kontext* ('während ich beim Arbeiten bin')
2. *Inhalt* ('die Anfrage enthält nur Interessen für Erwachsene')
3. *Zeit* ('zwischen 7:00 und 18:00')
4. *Reziprozität* ('wenn der Empfänger der Daten ähnliche Anfragen preisgibt')
5. *Anfrage-Ergebnis-Abhängigkeit* ('zeige die Anfrage, wenn der angeklickte Link auf eine Nachrichtenseite verweist')

Der Unterschied zwischen einer kontext- und einer inhaltsbezogenen Bedingung ist, dass Ersteres Bezug auf den Nutzer nimmt, wohingegen Zweiteres auf die verwendeten Wörter in der Anfrage abzielt. Das *Objekt* der Strategien kann hier entweder die Anfrage, der angeklickte Link oder beides sein. Die *Gruppe* spezifiziert Personen, die auf ein bestimmtes Objekt (nicht) zugreifen dürfen. Die Bedingungen, adressierte (Personen-) Gruppen und Objekte sind zueinander orthogonal, das heißt, wir haben beispielsweise keine Personengruppe identifiziert, die nur in einem bestimmten Kontext genannt worden ist. Aus diesem Grund evaluieren wir die folgenden Aspekte jeweils unabhängig voneinander.

Zwischenfazit: Strategien von Nutzern lassen sich auf eine einfache Struktur abbilden, bestehend aus Bedingungen, adressierten Personengruppen und Objekten, die es zu schützen gilt. Dies beantwortet unsere Forschungsfrage F3.

4.1.3.2. Pauschale Strategien

Die einfachste Variante einer Strategie erlaubt oder unterbindet die Preisgabe einer Anfrage, ohne Bedingungen oder Personengruppen zu definieren. Ein Beispiel ist 'GIB ZU KEINER ZEIT irgendetwas AN irgendjemanden preis'. 12 unserer 27 Teilnehmer

haben eine Strategie erstellt, die generell die Preisgabe aller Anfragen verbietet. 5 Teilnehmer haben eine Strategie definiert, die Anfragen an jeden weitergibt.

Zwischenfazit: Die Vielzahl der Strategien, die pauschal die Preisgabe von Informationen verbietet, ist konform zu Ergebnissen aus [AG05]. Das heißt, viele Personen sind moderat bis stark besorgt um ihre Privatheit. Ein PET muss also einen Mechanismus anbieten, mit dem die Privatheit eines Nutzers auf einfache Art konsequent vor jedem geschützt wird.

4.1.3.3. Bedingungen

Wir haben als nächstes die Bedingungen untersucht, die Teilnehmer in ihren Strategien definiert haben. Dies liefert auch die Antwort auf unsere Forschungsfrage F1. Von den Strategien mit einer Bedingung haben sich 39% (56) auf den Kontext bezogen, 11% (16) auf den Inhalt der Anfrage, 5% (7) auf Charakteristiken anderer Personen oder deren Strategien und 2% (3) auf zeitliche Einschränkungen.

Ein paar wenige Teilnehmer haben (auch nur maximal zwei) Bedingungen kombiniert, beispielsweise 'Wenn ich bei der Arbeit bin und die Zeit zwischen 7:00 und 18:00 ist', also eine kontextbezogene und eine zeitbezogene Bedingung. Die Teilnehmer haben keine Unterscheidungen zwischen den Anfragen selbst und mit ihnen verbundene Metadaten getroffen, wie zum Beispiel der Weitergabe des Namens der Person, die die Anfrage abgesetzt hat oder der Zeit.

Zwischenfazit: Im Allgemeinen definieren Nutzer in ihren Strategien relativ einfache Bedingungen. Die große Mehrheit der Nutzer definiert pro Strategie sogar nur eine Bedingung, jedoch mehrere Strategien.

Kontext Der Kontext beschreibt die Situation, in der die definierte Strategie Anwendung finden soll, zum Beispiel 'bei der Arbeit', 'bei der Planung eines Urlaubs' etc. Die Teilnehmer haben 39% (56) der 142 Strategien mit Bezug auf die Anfragen einem Kontext zugeordnet. Tabelle 4.2 gibt eine Übersicht über die Kontexte, die die Teilnehmer genannt haben. Für die verbleibenden 61% (86) der Strategien haben die Teilnehmer keinen bestimmten Kontext spezifiziert.

Zwischenfazit: Offensichtlich und erwähnenswert ist, dass Nutzer Strategien definieren, die von der Gesetzgebung nicht abgedeckt sind, beispielsweise durch die EU-Direktive [Par95]. Während der Gesetzgeber Kontexte definiert für den Informationsaustausch zwischen einer Person und einem Unternehmen, so definieren Nutzer Kontexte, die deutlich feingranularer vorgeben, zu welchem Zeitpunkt sie welche Informationen preisgeben.

Tabelle 4.2.: Nutzerkontexte der Strategien

Kontext	Häufigkeit
Zu Hause	21
Privates Surfen bei der Arbeit	10
Suche nach Erotikinhalten	5
Suche nach anbieterspezifischen Inhalten (z.B. 'youtube Videos')	5
Online einkaufen	4
Suche nach Krankheiten	3
Suche nach Arbeitsplätzen	2
Urlaubsplanung	2
Suche nach Partnervermittlungsseiten	1
Suche nach Personennamen	1
Suche nach Sportergebnissen	1
Suche nach Vermögensverwaltung	1

Inhaltsbedingungen Inhaltsbedingungen beziehen sich auf die einzelnen Terme einer Anfrage. Da der Inhalt einer Anfrage und der verwendetet Kontext, der den Nutzer motiviert hat die Anfrage abzusetzen, oftmals gleich sind, überlappen sich diese Bedingungen bis zu einem gewissen Grad. 11% (16) aller Strategien haben mindestens eine Bedingung ähnlich '*Wenn ich eine Anfrage absetze, die eines der Worte aus <Liste von Schlüsselworten> enthält, so zeige die Anfrage (nicht) <Liste von Personen(gruppen)>*' formuliert. Fünf Strategien beziehen sich auf den Anbieternamen, beispielsweise 'youtube', oder auf Nachrichtenseiten. Drei Strategen beziehen sich auf Personennamen, drei auf technische Begriffe und zwei auf Begriffe aus der Erotik. Tabelle 4.3 zeigt eine vollständige Liste aller formulierten Inhaltsbedingungen. Für die unteren drei Zeilen der Tabelle haben die Teilnehmer keine Schlagworte angegeben.

Eine eher überraschende Inhaltsbedingung bezieht sich auf die Struktur der Anfrage. Die Teilnehmer haben für 5 Strategien definiert, dass sie diese (nicht) preisgeben, wenn sie aus (weniger) mehr als einer bestimmte Anzahl von Termen besteht.

Zwischenfazit: Nutzer definieren Kontexte, die eine CSE automatisiert erkennen muss. Aus den Bedingungen bezüglich der Struktur folgern wir, dass Nutzer den Detailgrad der Preisgabe ihres Informationsbedürfnisses einschränken wollen. Das heißt, sie gehen davon aus, dass in längeren Anfragen detailliertere und damit die Privatheit stärker gefährdende Information gespeichert ist.

Tabelle 4.3.: Kategorien von Inhaltsbedingungen der Strategien

Inhalt	Häufigkeit	Beispiel
Anbietername	5	Youtube, FAZ
Personennamen	3	Name eines Experimentteilnehmers
Technik	3	XML, ABAP, PHP
Erotik	2	xxx
Schimpfwörter	1	xxx
Finanzen	1	Bank, Börse, Depot
Einkaufen	1	Kaufen, Shop
Verbotene Aktivitäten	1	download torrent
Krankheiten	1	–
Medikamente	1	–
Drogen	1	–

Zeitbedingung In drei Strategien haben unterschiedliche Nutzer Bedingungen definiert, die sich auf die Zeit und das Datum beziehen. Beispielsweise hat eine Strategie die Preisgabe von Suchinformationen während der Arbeitszeit untersagt. Die Begründung hier ist der Schutz vor Konkurrenten gewesen. Ein anderer Teilnehmer erlaubt die Preisgabe seiner Anfragen nur zwischen 6:00 und 20:00 Uhr. Eine Strategie hat sich auf das Datum bezogen und die Preisgabe in Zeiträumen von Feiertagen wie Weihnachten verboten (um keine Geschenkideen preiszugeben).

Zwischenfazit: Wir sind überrascht, dass Nutzer keinen Schutz vor Sequenzen von Anfragen berücksichtigen und keine zeitlichen Abstände zwischen dem Absetzen einer Anfrage und deren Weitergabe definieren. Anstelle dessen wollen sie eine Anfrage entweder ganz oder gar nicht preisgeben. Wie auch bei den ALWAYS (DO NOT)-Strategien sehen wir das als Hinweis, dass Nutzer ihre Strategien einfach halten wollen. Außerdem definieren sie Strategien, die Firmengeheimnisse schützen. Das geht weit über die klassischen Privatheitsaspekte hinaus.

Reziprozitätsbedingungen Bei Reziprozitätsbedingungen hängt die Entscheidung einer (Nicht-) Preisgabe von Information von Charakteristiken der Empfängerperson und deren Strategien beziehungsweise Verhalten ab. Drei Nutzer haben in Summe 7 Strategien dieser Art definiert. (i) Fünf von ihnen geben Anfragen nur dann preis, wenn der Empfänger selbst bereits ähnliche Anfragen abgesetzt hat. (ii) Ein Teilnehmer hat die Anfragen nur mit solchen Nutzern geteilt, die bereits vor ihm angemeldet waren. (iii) Ein Teilnehmer ist bereit, seine Daten dann weiterzugeben, wenn davon auszugehen ist, dass der Empfänger der Daten keine Rückschlüsse auf die Identität des Erstellers der Anfrage ziehen kann.

Zwischenfazit: Die Ergebnisse zeigen, dass einige Nutzer bereit sind, Information mit gleichgesinnten Personen zu teilen, etwa wenn sie an der gleichen Krankheit leiden. Darüber hinaus fordern Nutzer die Möglichkeit der Anonymisierung, das heißt ein Preisgeben der Anfragen nur dann, wenn die Identität geschützt bleibt. Reziprozitätsbedingungen zeigen außerdem, dass PETs aus dem Kontext des Web 1.0 nicht ohne Weiteres auf Web 2.0-Anwendungen übertragen werden können.

4.1.3.4. Personengruppen

Über Gruppen definieren Nutzer, welche Personen eine Anfrage (nicht) sehen dürfen. Wir haben uns dafür interessiert, welche Personengruppen wie 'Freunde' oder 'Familie' die Teilnehmer in ihren Strategien berücksichtigen. Eine grundlegende Erkenntnis ist, dass die adressierten Personengruppen in zwei Klassen zu unterteilen sind: (i) Gruppen, die nur Personen umfassen, die persönlich bekannt sind, und (ii) Gruppen mit Teilnehmern gleicher Charakteristik, die aber nicht persönlich bekannt sind. Beispielsweise umfasst die Gruppe 'Familienmitglieder' nur persönlich bekannte Personen, die ein Nutzer auch explizit aufzählen könnte. Kinder oder Männer hingegen haben die gleiche Eigenschaft das Alter beziehungsweise das Geschlecht betreffend, die Gruppe der Kinder oder Männer ist aber zu groß, um explizit aufgezählt werden zu können.

Unsere Teilnehmer haben 106 Strategien definiert, die sich auf unterschiedliche Personengruppen beziehen (Tabelle 4.4). 60% (82) der Strategien können wir der Klasse (i) zuordnen und 40% (42) der Klasse (ii). 83% (88) Strategien beziehen sich nur auf eine Personengruppe, 17% (18) Strategien beziehen mehrere Gruppen mit ein.

Die am häufigsten verwendete Gruppe aus Klasse (ii) sind Kollegen. Betrachtet man die Anfragen, bei denen die Teilnehmer solch eine Strategie anwenden wollen, ist das nicht überraschend. Anfragen, die keine Kollegen sehen sollen, sind zum Beispiel Suchanfragen nach neuen Jobs, Abteilungen etc. Bezogen auf Klasse (i) haben die Teilnehmer Freunde (33%) und Familie (18%) am häufigsten genannt. 20 von 27 (74%) Teilnehmern haben eine Strategie definiert, die das Teilen von Anfragen ausschließlich mit Freunden erlaubt.

Außerdem haben wir für jede Gruppe untersucht, ob ihr der Zugang zu den Anfragen erlaubt oder eher untersagt werden soll. So unterscheiden wir Strategien wie '*zeige meine Anfragen meinen Kollegen und Freunden*', die eine Weitergabe erlauben, und '*lass meinen Chef diese Anfrage nicht sehen*', Strategien die eine Preisgabe unterbinden. In unserer Studie haben die Teilnehmer beide Möglichkeiten häufig eingesetzt und haben sie auch vielfach kombiniert, wie '*Gib meinen Freunden Zugriff auf die Anfragen, nicht aber meiner Familie*'. 64% (68) unserer 106 Strategien mit Bezug zu einer Personengruppe ermöglichen den Zugriff, 36% (38) nutzen die Möglichkeit, Gruppen von dem Zugriff auszuschließen.

Tabelle 4.4.: Adressierte Personengruppen in den Strategien

Gruppe	Klasse	Häufigkeit	Zugriff erlaubt	Zugriff untersagt
Freunde	(i)	35	8	27
Familie	(i)	19	8	11
Bekannte	(i)	12	7	5
Freund/Freundin	(i)	4	1	3
Vorgesetzte	(i)	4	3	1
Ärzte	(i)	3	0	3
Lehrer	(i)	3	3	0
Eltern	(i)	1	0	1
Vermieter	(i)	1	1	0
Kollegen	(ii)	26	6	20
Kommilitonen	(ii)	6	2	4
Kinder	(ii)	3	3	0
Männer/Frauen	(ii)	3	0	3
Offizielle	(ii)	3	3	0
Mitbürger	(ii)	1	1	0

Zwischenfazit: Die Antwort auf unsere Forschungsfrage F2 ist, dass Nutzer besorgt darüber sind, was Freunde, Kollegen oder Familienmitglieder aus ihren Anfragen ableiten können. Das ist interessant, da es sich dabei genau um die sozialen Gruppen handelt, für die CSEs von besonderem Interesse sind [MH07]. Nutzer berücksichtigen in ihren Strategien Personengruppen, die (i) nur Personen umfassen, die persönlich bekannt sind, und (ii) Gruppen mit Teilnehmern gleicher Charakteristiken. Diese Unterscheidung ist für die Implementierung eines PETs wichtig: Die Personen aus Klasse (i) könnten gegebenenfalls automatisiert aus sozialen Netzwerkseite, dem E-Mailadressbuch oder aus einem Instant-Messenger-Kontaktverzeichnis erhoben werden. Dies ist deutlich schwieriger für Personengruppen der Klasse (ii). Außerdem muss es ein PET seinen Nutzern erlauben, sowohl positiv formulierte Strategien, die etwas erlauben, als auch negativ formulierte Strategien, die etwas verbieten, zu formulieren.

4.1.3.5. Strategien für Links

Unser Aufbau der Studie erlaubt eine Unterscheidung in die Objekte ‘Suchanfrage’ und ‘Angeklickter Link aus dem Suchergebnis’. Die 142 Strategien, die wir bisher analysiert haben, haben sich auf die Suchanfragen bezogen. Jetzt untersuchen wir die 105 Strategien, die sich auf Links beziehen, die nach einer Suche aus dem Suchergebnis angeklickt werden. 81 Strategien für Links sind Kopien der Strategien für die Suchanfragen. 7 neue Strategien wurden definiert. 17 Strategien stammen von Strategien für

Suchanfragen ab, wurden aber erweitert. Drei Teilnehmer haben keine Strategien für Links definiert.

Die 23% (24) der sich unterscheidenden Strategien können in drei Arten unterteilt werden: *Erstens* haben die Teilnehmer pauschale *ALWAYS DO NOT*-Strategien definiert. *Zweitens* sind Strategien weiter eingeschränkt worden, beispielsweise durch die Spezifikation zusätzlicher Gruppen, die einen Link (nicht) sehen dürfen. *Drittens* definieren einige Strategien Inhaltsbedingungen bezogen auf die URLs oder Beschreibungen der Webseiten hinter einem Link.

Interessant sind 6 Strategien mit einer Zeitbedingung: 3 fordern, dass Informationen über angeklickte Links nur für wenige Tage anderen Personen angezeigt werden. Solche Strategien erlauben den Einsatz von CSEs zum instantanen Informationsaustausch, verhindern aber die Erstellung von Langzeitprofilen. Drei Strategien erlauben nur den Austausch der letzten *n* Links. Vier Strategien definieren Anfrage-Ergebnis-Abhängigkeiten. Solche Abhängigkeiten sind eine neue Klasse von Bedingungen: Sie geben Anfragen abhängig von der Art des angeklickten Links weiter, zum Beispiel 'Gib keine *Anfragen* weiter, außer wenn der angeklickte *Link* ein Nachrichtenportal ist'.

Zwischenfazit: Unser Ergebnis bezüglich Forschungsfrage F4 ist, dass Nutzer für eine Vielzahl von Strategien nicht zwischen unterschiedlichen Objekten unterscheiden. Nennenswert ist, dass es Abhängigkeiten zwischen Objekten beziehungsweise den Privatheitspräferenzen für unterschiedliche Objekte geben kann.

4.1.4. Zusammenfassung und Diskussion

Mit unser Studie haben wir einen Einblick in ein wichtiges, neues Forschungsfeld erhalten: Die Privatheitspräferenzen von Nutzern bei der neuen Technologie der kollaborativen Suchmaschinen. Es ist unser Ziel gewesen, eine große Bandbreite an Strategien zu erheben, die insbesondere auf den Aussagen von Nutzern beruhen, die ein umfassendes Verständnis von CSEs haben. In unserem Fall mussten die Teilnehmer sogar einzelne Komponenten nachentwickeln, um dieses Verständnis sicherzustellen.

Durch die Auswertung von 247 Strategien haben wir erkannt, dass die Strategien auf eine einfache Struktur abgebildet werden können. Diese besteht aus einer oder mehreren Bedingungen und Objekten und bezieht sich auf unterschiedliche Personengruppen. Die Strategien beinhalten keine Bedingungen von mehr als zwei unterschiedlichen Typen. Die Teilnehmer haben sehr spezifische Strategien definiert, um zum Beispiel zu verhindern, dass Familienmitglieder Informationen über gekaufte Geschenke zu Weihnachten erhalten, ebenso wie sehr grobgranulare Strategien wie 'Gib unter keinen Umständen irgendetwas an irgendjemanden preis". Weiter beinhalten die Strategien Listen von sensiblen Schlüsselworten, auch wenn das Anbieten und die Pflege solcher Listen eine aufwändige Aufgabe ist. WordNet [Fel98] oder andere Thesauri könnten dies durch den Einsatz von Wortkonzepten vereinfachen. Die Strategien bezie-

hen sich außerdem auf unterschiedliche Personengruppen. Einige von diesen Gruppen entsprechen, wenn auch nur sehr grob, den Kontakten, die Nutzer in Form von sozialen Netzwerkseiten und den darin enthaltenen Freundeslisten bereits explizit gemacht haben. Ein Ansatz, die Personen der sozialen Gruppen abzuleiten, ist der Einsatz von Schnittstellen zu den sozialen Netzwerkseiten, wie beispielsweise OpenSocial[2], oder die Anbindung von Messenger Daten [LH08]. Schwieriger gestaltet sich die Identifikation der Teilnehmer von Gruppen, bei denen die Teilnehmer nicht explizit aufgezählt werden können. Beispiele sind 'Männer' oder 'Kinder', insbesondere wenn man zu deren Definition weitere, für den Dienst eigentlich unnötige Daten wie das Alter erheben muss. Darüber hinaus adressieren die Strategien Zeit- und Inhaltsbedingungen und unterscheiden zwischen Anfragen und angeklickten Links aus dem Suchergebnis. Unsere Ergebnisse zeigen auch einen Unterschied zwischen Anforderungen an PETs im Web 1.0 zu Anforderungen an PETs im Web 2.0. So fordern Nutzer zum Beispiel reziproke Strategien, die Charakteristiken, Strategien und das Verhalten anderer Nutzer miteinbeziehen. Dies zu ermöglichen ist eine Herausforderung, da bei konfligierenden Strategien der Informationsaustausch vollständig zum Erliegen kommen kann. Beispielsweise könnten die Strategien zweier Nutzer jeweils darauf warten, dass der andere Nutzer etwas zuerst preisgibt. Zusätzlich kann die strategische Definition von reziproken Strategien dazu führen, dass ein potentieller Angreifer die Privatheitspräferenzen anderer Nutzer erschließen kann. Nutzer fordern Anonymität, das heißt, eine Information wird nur dann weitergegeben, wenn der Ersteller nicht identifiziert werden kann. Das ist ebenfalls schwierig umzusetzen, beispielsweise wenn ein Nutzer einen Term nur anonymisiert preisgeben möchte, ein anderer jedoch konkret. Unter Umständen würde das zu Wechselwirkungen zwischen den Strategien führen, so dass die Information der Person, die diese personalisiert preisgeben will, ebenfalls anonymisiert werden muss. Nicht zuletzt ist eine Anonymisierung von Suchanfragen schwierig – auch wenn alle Nutzer die gleiche Strategie anwenden (siehe dazu Kapitel 6).

Da wir im Vorfeld existierende Mechanismen zum Schutz von Privatheit und Bedrohungen der Privatheit mit den Teilnehmern unserer Studie nicht thematisiert haben – dies hätte mit an Sicherheit grenzender Wahrscheinlichkeit zu einer Strategie geführt, die genau vor den vorgestellten Bedrohungen schützt – ist die Menge der Strategien möglicherweise unvollständig. Die Studie zeigt jedoch das Spektrum der Strategien auf, die Nutzer zum Schutz ihrer Privatheit sinnvoll empfinden. Das heißt, wir haben festgehalten, was ein Privatheitsmechanismus aus Sicht seiner Nutzer mindestens umfassen sollte. Wir gehen davon aus, dass die Berücksichtigung dieser Anforderungen bei der Entwicklung von PETs das Vertrauen der Nutzer in einen Dienstgeber steigern wird.

[2] `http://code.google.com/apis/opensocial/`

4.2. PETs und Kollaboration bei standortbezogenen Diensten

4.2.1. Privatheitsprobleme standortbezogener Dienste

Standortbezogene Dienste (Location Based Services, LBS) sind ein wesentlicher Bestandteil vieler Anwendungsdomänen geworden. Google Latitude[3] ermöglicht beispielsweise den Austausch seiner Positionsdaten mit Freunden und Bekannten, [Tim08] das Finden von Freunden abhängig von der eigenen Position und Panoramio[4] das Annotieren von Orten mit Fotos.

Die Nutzung standortbezogener Dienste erfordert von den Nutzern die Preisgabe ihrer Position, das heißt sensibler, privater Information.

Beispiel 11: Eine Person nutzt eine Anwendung zum Verschlagworten von Orten (Geo-Tagging). Die aktuelle Position wird per Global Positioning System (GPS) festgestellt. Dies erlaubt es anderen Personen anhand generierter Schlagworte herauszufinden, (1) wo die Person zu einem bestimmten Zeitpunkt gewesen ist, (2) welche Daten sie interessant findet und (3) ihr Bewegungsprofil.

Das Bewegungsprofil kann dabei entweder durch die Verbindung getaggter Orte erfolgen oder, wie etwa bei Google Latitude, durch die Übermittlung hochauflösender Bewegungsdaten. Letzteres erlaubt weitere Rückschlüsse, zum Beispiel auf die Geschwindigkeit und damit auf die Fitness eines Radfahrers oder Joggers. Die Forschungsliteratur hat eine Reihe von 'Privacy-Enhancing Technologies' (PET) für standortbezogene Dienste vorgestellt. Diese unterscheiden sich in ihrer Komplexität, dem vom Nutzer geforderten Bewusstsein sowie des vom Nutzer geforderten Verständnisses von LBS. Es ist jedoch eine offene Frage, welche dieser PETs auch tatsächlich im Kontext von LBS so eingesetzt beziehungsweise so benutzt werden können, dass Nutzer mit den PETs ihre Privatheitspräferenzen durchsetzen können.

Wir haben mit Hilfe einer Nutzerstudie [7] die folgenden Forschungsfragen untersucht:

F1: Welche Informationen geben Personen im LBS Kontext preis? Um PETs zu etablieren ist es wichtig zu wissen, (1) welche Daten LBS Nutzer preisgeben und (2) welche Daten sie nur für die private Nutzung generieren. Wir unterscheiden zwischen der Positionsinformation zu einzelnen Orten, aus Ortsangaben resultierende Spuren (Tracks), den Inhalt, zum Beispiel die Annotationen an Orten, und Metadaten, zum Beispiel den Zeitpunkt der Erstellung des Inhalts. Wir wollen herausfinden, wie viele Daten Nutzer erstellen und um welche Art von Daten es sich handelt. Außerdem interessiert uns, ob die preisgegebenen Daten die Alltagsmittelpunkte der Anwender widerspiegeln.

F2: Welche sozialen Gruppen dürfen welche private Daten einsehen? Standortbe-

[3] www.google.com/latitude, April 2010

[4] www.panoramio.com, April 2010

zogene Dienste werden häufig zum Austausch von Positionsdaten genutzt, zum Beispiel um sich an einem bestimmten Ort zu verabreden oder Freunde zu finden. Wir möchten beobachten, welche Art von Daten Personen allen (public) zur Verfügung stellen und welche nur ihren direkten Freunden und Kontakten.

F3: Wie nutzen Anwender unterschiedliche PETs? Sind die Anwender nicht in der Lage, ein PET korrekt einzusetzen, wird es in der Praxis versagen. Wir untersuchen, wie Anwender mit unterschiedlichen PETs umgehen. Die untersuchten PETs umfassen dabei ganz einfache Mechanismen, wie einen Schalter für die GPS Übertragung, sowie komplexe Mechanismen (siehe auch Kapitel 2.4). Alle untersuchten PETs unterscheiden sich weiter in dem Grad des erforderlichen Bewusstseins, der Aufmerksamkeit und dem Verständnis zur korrekten Nutzung.

F4: Welche PETs bevorzugen Anwender? PETs müssen die Erwartungen und Wünsche der Nutzer erfüllen. Wir werden Anwender PETs anhand ihrer Erfahrungen bezüglich der Brauchbarkeit der PETs im alltäglichen Einsatz bewerten lassen.

Diese Forschungsfragen lassen sich nur mittels einer Nutzerstudie beantworten. Dies ist jedoch schwierig: *Erstens* sind Privatheitspräferenzen von Personen, die in simulierten Umgebungen (offline), Fragebögen oder in Laborexperimenten erhoben werden, nicht notwendigerweise die, welche die Anwender auch in der Realität erkennen lassen [SGB01]. So neigen Personen dazu, ihren Bedarf an Privatheit überhöht anzugeben. *Zweitens* erhalten Mobiltelefone mit GPS, einer breitbandigen Datenanbindung (zum Beispiel Universal Mobile Telecommunications System, UMTS) und permanenter Anbindung an das Datennetz (engl. flatrate), gerade erst Einzug in den Massenmarkt. Kunden können sich der mit den Technologien und verbundenen Diensten einhergehenden Bedrohungen noch gar nicht bewusst sein. Darüber hinaus können wir nicht davon ausgehen, dass der normale Anwender Vorschläge zu PETs aus der Forschungsliteratur kennt [GG03, KYS05, GL04, MCA06]. *Drittens* erfordert es einen sehr hohen Einsatz und große finanzielle Aufwendungen, Personen mit neuster Technik auszustatten. Außerdem muss eine Studie so gestaltet sein, dass technische Randbedingungen das Experiment nicht beeinflussen und die Teilnehmer der Studie über einen längeren Zeitraum motiviert bleiben. *Viertens* resultieren viele Privatheitsbedrohungen aus dem Austausch beziehungsweise Teilen (engl. sharing) von Daten zwischen Nutzern und sozialen Interaktionen. Folglich können aussagekräftige Ergebnisse nur dann gewonnen werden, wenn wir eine Gruppe von Teilnehmern finden, die schon vor dem Experiment sensible Informationen miteinander ausgetauscht hat.

Wir haben eine umfangreiche Nutzerstudie durchgeführt, die die vier zuvor beschriebenen Herausforderungen berücksichtigt. Unser primäres Interesse ist die Privatheitsbeziehung zwischen Nutzern. Das heißt, wir betrachten den Anbieter des Dienstes als vertrauenswürdige Instanz und vernachlässigen Privatheitsaspekte zwischen Anbietern und Dienstnutzern (eine umfangreiche Betrachtung dieses Aspektes befindet sich in Kapitel 5.1 und Kapitel 5.2). Wir haben eine mobile Anwendung zum Taggen von Orten entwickelt. Tagging ist eine beliebte LBS-Anwendung, die von den Nutzern ge-

nerierte standortbezogene Inhalte (engl. tags) zwischen vielen Nutzern austauscht. Da Tags eine relativ generische Form des Inhaltes sind, gehen wir davon aus, dass unsere Ergebnisse auch für andere Formen standortbezogener Dienste gültig sind.

Wir sind die ersten, die Privatheitmechanismen für LBS in einer Studie mit realen Nutzungsbedingungen untersuchen. In unserer Studie haben die Teilnehmer unsere Geo-Tagging-Anwendung für zwei Wochen von einem XDA aus in ihrem Alltag verwendet. Außerdem haben wir eine Web-Applikation erstellt, die es den Teilnehmern erlaubt, am Computer eine Übersicht aller Tags auf einer Karte von Google Maps anzuzeigen. Ohne unser Interesse am Thema Privatheit preiszugeben, haben wir die Teilnehmer aufgefordert, interessante Orte zu taggen, wie 'bestes Café' oder 'Haus eines Freundes' etc. Um eine reale Bedrohung für die Teilnehmer zu schaffen, haben wir Freunde, Familienmitglieder, Lehrer, Klassenkameraden und Bekannte eingeladen, an dem Experiment teilzunehmen. Diese Gruppen können die Tags, Tracks und Metadaten einsehen, die die Teilnehmer preisgegeben haben; natürlich nur, wenn der Ersteller der Daten kein PET genutzt hat, um die Inhalte privat zu machen. Um die Privatheitsbedürfnisse der Teilnehmer, den Umgang mit Privatheitsbedrohungen sowie den Umgang mit PETs zu untersuchen, haben wir einige wohlbekannte PETs implementiert.

Unsere Analyse zeigt, dass die Teilnehmer bei der Preisgabe von Daten kaum Unterscheidungen zwischen sozialen Gruppen machen. Für mehr als 81% aller Daten haben die Teilnehmer Informationen entweder für alle oder niemanden sichtbar gemacht. Andersrum betrachtet haben die Teilnehmer in 17% der Fälle die Möglichkeit genutzt, feingranulare Privatheitseinstellungen vorzunehmen. Das bedeutet, die Teilnehmer sind sich der Privatheitsbedrohung in vielen Fällen durchaus bewusst. Wir haben außerdem festgestellt, dass die Teilnehmer unterschiedlich komplexe PETs nutzen wollen. Erstaunt haben wir weiter beobachten können, dass sich Teilnehmer über die Privatheit der Freunde mehr Sorgen machen als über ihre eigene.

4.2.2. Szenario, Informationsfluss und implementierte PETs

In diesem Abschnitt beschreiben wir das Szenario unserer Studie (Abschnitt 4.2.2.1), bei LBS im Allgemeinen und unserer Studie im Speziellen ausgetauschte Daten (Abschnitt 4.2.2.2) und PETs für den Einsatz in LBS (Abschnitt 4.2.2.3).

4.2.2.1. Mobiles Geo-Tagging

Der vielleicht bekannteste standortbezogene Dienst ist das Geo-Tagging, das heißt das Verschlagworten von Orten. Der Dienst Panoramio erlaubt das Annotieren von Orten mit Fotos anstelle von Worten. Nutzer können diese Bilder anhand eines Google-Maps-Mashups durchsuchen, mit zusätzlichen Tags versehen, kommentieren oder die Kamerainformation und Daten zum Fotografen abrufen. Wie man leicht erkennen kann, erlaubt Panoramio die Verknüpfung von Orten, an denen sich eine Person aufgehal-

ten hat, mit der von der Person an diesem Ort preisgegebenen Information. Dies lässt wiederum Rückschlüsse auf die Intention und die Interessen des Nutzers zu. Um zu beobachten, wie Leute mit Privatheitsbedrohungen im LBS-Kontext umgehen, haben wir einen Dienst zum Verschlagworten von Orten entworfen. Die technische Grundlage liefert dabei der LBS-Framework Streamspin [WJPT07] und eine Oracle Datenbank mit einer Erweiterung für räumliche Daten (engl. spatial extension). Unser LBS speichert den realen Namen der Person, die ein Tag an einem Ort anlegt, sein Pseudonym, die Zeit der Erstellung des Tags, die Geo-Koordinaten und Tracks als Kombination aufeinanderfolgender Orte. Die Nutzer unseres LBS können sich alle Tags und zugehörigen Metadaten, die der Ersteller eines Tags sichtbar gemacht hat, ansehen und danach suchen. Unsere Anwendung kann sowohl von einem XDA (Modell HTC Trinity) aus, als auch von einer webbasierten Applikation auf jedem Rechner genutzt werden. Mit Hilfe des XDAs kann ein Nutzer Tags bezogen auf seine aktuelle Position erstellen. Das heißt, unsere mobile Anwendung auf dem XDA verbindet sich mit unserem zentral laufenden LBS und speichert für jeden Tag ein Tupel `<user, timestamp, location, tag list>` in der Datenbank. Außerdem erlaubt die mobile Anwendung auf dem XDA das Suchen nach Tags anhand einer textuellen sowie einer Umkreissuche. Die auf dem mobilen Endgerät angezeigten Tags aktualisieren sich in Echtzeit. So werden, wann immer der Nutzer seine Position wechselt, ihm die Tags angezeigt, die in seinem aktuellen Umfeld angelegt worden sind. Die Web-Applikation zeigt die Tags, Orte und Tracks auf einem Google-Maps-Mashup an. Das Mashup erlaubt es den Nutzern auch, mit dem XDA preisgegebene Daten zu bearbeiten. Um den Teilnehmern bereits initial interessante Informationen auf dem Handy anzeigen zu können, haben wir Tags zu 200.000 mit Geo-Referenzen versehenen Wikipediaartikeln extrahiert.

4.2.2.2. Datenpreisgabe bei LBS

Eins unserer Ziele ist es, die Privatheitsbedrohungen beim Austausch von Informationen zwischen Personen im LBS-Kontext zu untersuchen. Dazu betrachten wir die folgenden vier Datentypen, die Nutzer preisgeben und die zu unterschiedlichen Arten von Bedrohungen führen:

Inhalt Inhalt bezeichnet die Informationen, die Personen explizit generieren [AN07]. Beispielsweise sind die Inhalte einer standortbezogenen Tagging-Anwendung die Tags oder als solche verwendete Fotos, die ein Nutzer des LBS an seinem aktuellen Ort erstellt. Es ist klar, dass diese Inhalte nicht für jeden bestimmt sein müssen.

Position Mobile standortbezogene Dienste erheben die aktuelle Position ihrer Nutzer wann immer diese Inhalte generieren oder Inhalte in ihrem Umfeld anfragen [MCA06]. Manche LBS verarbeiten auch Informationen zu Orten, an denen ein Nutzer zuvor schon einmal gewesen ist. Informationen über die Orte, die ein Nutzer besucht hat, können privat sein. Ein offensichtliches Beispiel ist das

eigene Haus aber auch das gesamte Netzwerk besuchter Freunde.

Tracks Tracks und daraus resultierende Bewegungsprofile entstehen, wenn ein LBS die Bewegungen seiner Nutzer kontinuierlich aufzeichnet. Beispielsweise könnte ein Nutzer kontinuierlich seine aktuelle Position an den Anbieter senden, um von diesem über Freunde im direkten Umfeld informiert zu werden. Da diese Anfragen die aktuelle Position des Nutzers beinhalten müssen, kann der Anbieter die Bewegungen des Nutzer detailliert nachvollziehen, herausfinden, wo dieser wohnt, Gewohnheiten und Beziehungen ableiten, etc.

Metadaten Metadaten beziehen sich auf den Inhalt, die Position und die Tracks. Metadaten umfassen den Namen des Nutzers, dessen Pseudonym, die Zeit der Erstellung eines Tags, etc. Die Art der Metadaten ist abhängig vom Kontext. Fotos beinhalten oftmals Exif[5] Informationen über das Kameramodell, Objektiv, Brennweite etc. Metadaten können spezifisch für eine Person und damit identifizierend sein, das heißt, eine Bedrohung für die Privatheit darstellen.

4.2.2.3. Technologien zum Schutz der Privatheit im LBS Kontext

Eine Vielzahl von PETs für LBS sind verfügbar. Diese reichen von einfachen, intuitiven Mechanismen bis hin zu ausgefeilten Systemen, die ein grundlegendes Verständnis möglicher Privatheitsbedrohungen und des technischen Hintergrunds erfordern. Um eine breite Spanne unterschiedlicher Ansätze zu untersuchen, haben wir uns für die Implementierung von fünf PETs, jedes stellvertretend für eine Klasse von PETs, entschieden. Die PET-Klassen haben wir anhand der Forschungsliteratur und der Betrachtung der (wenigen) realen Dienste, die PETs anbieten, identifiziert. Wir unterscheiden PETs für Tags und die für Tracks. Für Tags und Tracks gibt es jeweils zwei Klassen: einfache, pauschal wirkende PETs und PETs, die feingranulare Einstellungen erlauben. Für Tracks untersuchen wir zusätzlich ein komplexes, nur schwer verständliches aber automatisiert arbeitendes PET, das auf Anonymisierung basiert (siehe dazu auch Kapitel 2.4.2).

Da wir den Austausch von sensiblen Daten zwischen unterschiedlichen Personengruppen untersuchen möchten, beinhaltet die Auswahl der PETs auch solche, die eine Unterscheidung der Privatheitseinstellungen für unterschiedliche Gruppen erlauben. Zu diesem Zweck haben die Teilnehmer Personen benannt und den Kategorien *Lehrer, Eltern, Klassenkameraden, Freunde* und *Bekannte* zugeordnet. Die benannten Personen haben ebenfalls Zugang zu dem System erhalten, um eine reale Bedrohung für die Privatheit der Teilnehmer zu generieren.

Als nächstes stellen wir die implementierten PETs vor. Die Teilnehmer können zwei PETs nutzen, um ihre Präferenzen für Tags zu definieren:

[5] engl. Exchangeable Image File Format. Ein Standard zur Speicherung von Metadaten zu Bildern.

$PET_{checkbox}$ **(Öffentlich/Privat Checkbox)** Dieses PET ist das einfachste betrachtete PET und zugleich intuitiv bedienbar. Es handelt sich dabei um eine einfache Checkbox, das heißt das Setzen eines Häkchens, über das der Nutzer entscheidet, ob eine Information entweder privat ist oder allgemein zugänglich gemacht werden soll.

PET_{fine} **(Feingranulare Einstellung)** Mit Hilfe dieses PETs kann ein Nutzer für jeden Typ von Daten seine Privatheitseinstellungen individuell vornehmen. Wir haben dieses PET in unsere Web-Applikation integriert. Nachdem der Nutzer seine Tags definiert hat, kann er über die Webseite eine Feineinstellung seiner Präferenzen vornehmen. Mit Hilfe von Auswahlboxen können die Nutzer für jede Position, jeden Tag, den Namen, das Pseudonym und den Zeitpunkt der Erstellung sowie für jede soziale Gruppe festlegen, wer was sehen darf. Ein Nutzer kann so beispielsweise seine Präferenz "Mache das Tag 'meine Schule' zugänglich für meine Freunde, aber verstecke meinen Namen und den Zeitpunkt" umsetzen (eine umfangreiche Betrachtung von Strategien aus einem anderen Web 2.0-Kontext liefert Abschnitt 4.1). Wir haben sichergestellt, dass es den gleichen Aufwand bedarf ein Tag privat oder öffentlich zu machen.

Um die Privatheitspräferenzen für Tracks beobachten zu können, haben wir die folgenden drei PETs implementiert. Darüber hinaus haben wir PET_{fine} so angepasst, dass es sich auch auf Abschnitte von Tracks anwenden lässt.

PET_{areas} **(Private Bereiche)** Mit Hilfe dieses PETs können die Teilnehmer Bereiche definieren, innerhalb derer ihre Tracks niemandem zur Verfügung gestellt werden. Somit kann ein Nutzer, der die Bewegungen einer anderen Person beobachtet, diese nur bis an die Grenze des privaten Bereichs verfolgen. Die Definition der Bereiche erfolgt über ein Google-Maps-Mashup. Das heißt, der Nutzer kann (geschlossene) Polygone zeichnen, innerhalb derer er geschützt sein möchte. Beispiele solcher Polygone reichen von einfachen Rechtecken um die eigene Wohnung zu Polygonen, die die ganze Welt umfassen. Private Bereiche werden vorab am Rechner definiert. Der Abgleich, ob ein Bereich betreten wird oder nicht, findet auf dem XDA statt, zeitlich somit vor der Datenerhebung durch den Anbieter.

PET_{switch} **(GPS Schalter)** Dieser Mechanismus ist durch die Arbeit von [BD03] motiviert. Er stellt die einfache Option dar, das GPS Modul ein- beziehungsweise auszuschalten. Es verhält sich dabei ähnlich dem PET, das auch bei Google Latitude implementiert ist. Es erlaubt dem Nutzer die direkte und unmittelbare Entscheidung über die Preisgabe seiner Position. Auf der anderen Seite bedarf es andauernder Aufmerksamkeit durch den Nutzer, um die Einstellung des GPS-Schalters im Auge zu behalten.

PET_{anon} **(Anonymisierer)** Dieses PET, motiviert durch [MCA06], verwendet die Notation von k-Anonymität so, dass Anfragen nach Tags im Umfeld nicht einer einzelnen Person zugeordnet werden können. Anders ausgedrückt ist jeder Nutzer und dessen Anfragen nicht von k-1 anderen Nutzern unterscheidbar. Der Anonymisierer funktioniert wie folgt: Bevor eine Anfrage abgesetzt wird, spezifiziert der Nutzer ein k. Dieses k repräsentiert die Anzahl der Nutzer, von denen er nicht unterscheidbar sein möchte. Man stelle sich einen Nutzer vor, der eine Anfrage nach Tags in seiner Umgebung stellt. Der Anonymisierer erweitert die Anfrageregion nun so lange, bis theoretisch auch k-1 andere Personen diese Anfrage hätten absetzen können. Der LBS antwortet mit den Tags für die erweiterte Region, welche dann lokal auf dem XDA auf die Parameter der ursprünglichen Anfrage gefiltert werden. Es ist bekannt, dass dieser Algorithmus in zwei Fällen nicht vor der Erstellung detaillierter Tracks schützt: (i) bei häufigen oder kontinuierlichen Anfragen (aus diesem Grund mussten die Anbieter für dieses PET die Ansicht manuell aktualisieren) und (ii) in Regionen mit einer hohen Nutzerdichte. Bei einer hohen Nutzerdichte muss die angefragte Region, damit sie k Nutzer umfasst, nur unwesentlich vergrößert werden. Ein Beobachter kann also auf die Position des Anfragestellers rückschließen. Aus diesem Grund schlägt die Literatur, zum Beispiel auch [MCA06], die Einführung einer minimalen Größe des angefragten Bereiches und einen lokalen, adaptiven Optimierer für die Bereiche vor. Andere schlagen die Gruppierung ähnlicher Tracks vor. Wir haben uns jedoch entschieden, den Anwender nur den Parameter k einstellen zu lassen, um ihn nicht zu überfordern. Die Nutzung dieses PETs erfordert ein grundlegendes Verständnis seiner Funktionsweise und der Privatheitsbedrohungen bei LBS.

4.2.3. Methodik der Nutzerstudie

In diesem Abschnitt beschreiben wir die zentralen Entwurfsentscheidungen unserer Studie sowie deren Aufbau und Durchführung.

4.2.3.1. Zentrale Entwurfsentscheidungen

Für die Studie haben wir fünf zentrale Entwurfsentscheidungen getroffen:

Intensive Studie Wir haben uns für die Durchführung einer Studie entschieden, die eine starke Einbindung der Teilnehmer erfordert, die unter realen Bedingungen und mit einer relativ kleinen Anzahl von Teilnehmern durchgeführt werden soll. Der Grund dafür ist folgender: (1) Um zu aussagekräftigen Ergebnissen zu kommen, ist Kompetenz bezüglich der verwendeten Technologie erforderlich [Bab07]. Deshalb müssen wir die Teilnehmer beim Einsatz von GPS, den XDAs, und unserer Anwendung *intensiv trainieren*. (2) Teilnehmer von Offline-Studien tendieren dazu, ihre Privatheitsanforderungen zu überschätzen [SGB01]. Wir gehen davon aus, dass wir realistische Ergebnisse nur durch den Einsatz

unseres LBS in der *Realwelt* und durch die Integration der Anwendung in den *Alltag* der Teilnehmer erreichen. Zur Evaluierung unserer Forschungsfragen beobachten wir das Verhalten der Teilnehmer.

Zweiteilung des Experiments Wir haben die Durchführung des Experiments, das heißt die Evaluierung der PETs, in zwei Phasen unterteilt: (i) eine Tagging-Phase, in der wir uns auf die Positionen, Inhalte und die Metadaten der Tags konzentrieren, und (ii) die Trackaufzeichnungsphase, in der wir uns der Analyse von Bewegungsprofilen widmen. Da einige PETs zum Schutz der Tracks miteinander in Konflikt stehen – zum Beispiel kann man den Effekt der geschützten Bereiche (PET_{areas}) nicht messen, wenn das GPS abgeschaltet ist (GPS_{switch}) – untersuchen wir in der zweiten Phase die PETs nacheinander und voneinander isoliert. Außerdem möchten wir dadurch vermeiden, dass wir die Teilnehmer mit der Kombination teilweise komplexer PETs überfordern.

Web-Applikation Wir haben aus zwei Gründen Teile unserer Anwendung als Web-Applikation entworfen und implementiert: Erstens benötigen einige PETs eine detaillierte Parametrisierung. Damit unintuitive Nutzerschnittstellen nicht die Ergebnisse unserer Studie beeinflussen, haben wir den Dialog zur Konfiguration der Nutzerpräferenzen für PET_{fine} und PET_{areas} als Web-Applikation ausgestaltet. Zweitens haben wir wie beschrieben Freunde, Bekannte etc. aus dem Umfeld der Teilnehmer akquiriert. Diese Personen haben über die Web-Applikation die Möglichkeit, die preisgegebenen Informationen einzusehen.

Teilnehmer Typischerweise werden Tagging-Dienste zur Unterstützung des Auffindens von Informationen, zum Kundtun einer Meinung, zum Austausch von Informationen oder zu spielerischen sowie Wettbewerbszwecken eingesetzt[MNBD06]. Somit brauchen wir für unsere Studie technologieaffine Personen. Zugleich müssen die Teilnehmer Teil des gleichen sozialen Netzwerkes sein und nah beieinander wohnen, um die Tags von Freunden auch wirklich wahrzunehmen. Außerdem sollen sie intrinsisch motiviert sein, sich miteinander zu treffen. Nur so können wir die Effekte der PETs auf die sozialen Netzwerke der Teilnehmer messen. Nicht zuletzt sollen unserer Teilnehmer eine interessante Zielgruppe für die Anbieter von LBS sein. Wir haben uns für eine Gymnasialklasse (11. Klasse) mit 25 Schülern zwischen 16 und 17 Jahren entschieden. 10 Freiwillige (acht weiblich, zwei männlich) von ihnen haben wir mir XDAs ausgerüstet. Um reale Privatheitsbedrohungen vorzufinden, haben wir die Eltern der Teilnehmer, deren Lehrer und andere Schüler, die keinen XDA erhalten haben, mit einem Zugang zu der Web-Applikation ausgerüstet. Folglich können viele Personen aus dem sozialen Umfeld der Teilnehmer die generierten Inhalte, Metadaten und Tracks, die die Teilnehmer sichtbar gemacht haben, einsehen. Die Teilnehmer mit XDA haben alle Personen des Experiments den sozialen Gruppen zugeordnet, beispielsweise hat ‘Tochter Schmidt’ ‘Vater Schmidt’ in die Gruppe ‘Eltern’ aufgenommen. Darüber hinaus haben wir Daten in anderen Orten simuliert und den Teilneh-

mern mitgeteilt, dass noch weitere Schulen und Firmen das gleiche Experiment machen. So haben sie davon ausgehen müssen, dass das öffentliche Preisgeben von Informationen private Inhalte auch völlig Unbekannten zugänglich macht.

Anreize Um die Teilnehmer zur Nutzung unseres LBS zu motivieren, durften sie für jeweils fünf getaggte Orte eine Frei-SMS verschicken. Ebenso haben sie für jede halbe Stunde, in der sie die Anwendung angeschaltet haben, eine Frei-SMS erhalten. Diese Auszahlung ist unabhängig von dem Ort, an dem die Teilnehmer sich befinden, und von den preisgegebenen Inhalten erfolgt. Das heißt, wir haben die Präferenzen durch unseren Mechanismus nicht beeinflusst, da die Auszahlung für ein privates Tag gleich der für ein öffentliches Tag ist. Vorwegnehmend können wir sagen, dass unsere Teilnehmer die XDAs und unseren LBS intensiv genutzt haben.

4.2.3.2. Aufbau und Durchführung

Wir haben die Studie in vier Phasen unterteilt: einer *Einführungs-*, *Tagging-*, *Trackaufzeichnungs-* und *Abschluss*phase. Jede Phase hat mit einem gemeinsamen Treffen gestartet und geendet. Dabei haben wir die Teilnehmer entsprechend der Phase instruiert und mit Hilfe von Kontrollfragen in Fragebögen das Verständnis der Teilnehmer zu dem Einsatz der PETs kontrolliert. Während des gesamten Experiments haben wir nahezu jeden Tastendruck und jede vom XDA ermittelte GPS Position gespeichert. Das bedeutet, dass unsere Protokolle die kompletten Bewegungsprofile der Teilnehmer während des Experimentzeitraumes umfassen. Das gilt auch für solche Zeitpunkte, in denen ein Teilnehmer der Ansicht war, dass das GPS ausgeschaltet sei[6].

Eine Beschreibung der Realisierung der Tagging-Anwendung und Screenshots befinden sich im Anhang B.

Einführung Um eine plausible Motivation für unsere Anwendung zu geben, ohne dabei jedoch gleichzeitig unser Interesse an Privatheit preiszugeben, haben wir zuerst beliebte Tagging-Dienste und LBS wie del.icio.us und mobile Stadtführer vorgestellt. Mit einem ersten Fragebogen haben wir Informationen über (i) demographische Daten der Teilnehmer, (ii) deren Nutzungsgewohnheiten betreffend Mobiltelefonen und Internetdiensten eingeholt sowie (iii) allgemeine Fragen zum Thema Privatheit gestellt. Ziel von (iii) ist der Ausschluss von beispielsweise Datenschutzfundamentalisten [ACR99].

[6] Wir haben die Lehrer der Schüler von Anfang an über unser primäres Interesse an dem Thema Privatheit und dem Umgang mit PETs informiert. Außerdem haben wir keinerlei Daten Dritten zugänglich gemacht, wenn die Teilnehmer im Gerät den Status 'Privat' ausgewählt haben. Es ist jedoch nur auf diesem Wege möglich, nicht sein Interesse an dem Thema Privatheit preiszugeben und den wirklichen Unterschied zwischen realen Privatheitspräferenzen und den behaupteten zu messen. Weiter haben von allen Teilnehmern die Eltern eingewilligt, dass ihre Kinder an dem Test der Tagging-Anwendung teilnehmen. Wir haben, alleine schon aus Gründen der Vergleichbarkeit, keine Daten erhoben, die nicht jeder vergleichbare LBS auch erhebt.

Außerdem haben wir unser XDA mit der mobilen Tagging-Anwendung vorgestellt. Wir haben die Funktionalität unserer Anwendung erklärt und mit den Teilnehmern gemeinsam auf dem Universitätscampus Trainingsobjekte getaggt. Anschließend haben wir jedem Teilnehmer gezeigt, wie diese Trainingsobjekte auf der Webseite angezeigt werden. Zuletzt haben wir den Teilnehmern gesagt, dass Dienste zum standortbezogenen Tagging zumeist Möglichkeiten bieten einzustellen, wer welche Inhalte sehen darf. Zu diesem Zweck haben wir die beiden Privatheitsmechanismen $PET_{checkbox}$ und PET_{fine} eingeführt. Um eine realistische Privatheitsbedrohung zu generieren, haben die Teilnehmer anschließend ihre jeweiligen Bekannten, Freunde, den Lehrer und Verwandte benannt und in dem System eingetragen.

Tagging Die Tagging-Phase hat eine Woche gedauert. Wir haben unsere Teilnehmer motiviert, das XDA in ihrem Alltag einzusetzen und unterschiedliche Dinge zu taggen. Wir haben Beispiele wie Sehenswürdigkeiten, Treffpunkte, das eigene Haus und das Haus von Freunden vorgegeben. Im Hintergrund haben wir jede Bewegung des XDAs aufgezeichnet. Am Ende der Woche haben wir unser Interesse am Thema Privatheit im LBS Kontext preisgegeben. Zuerst haben wir einen Fragebogen zur Nutzbarkeit unserer Tagging-Anwendung, zu technischen Problemen, die unsere Ergebnisse bis zu diesem Punkt beeinflusst haben könnten, und zu dem Anreizmechanismus gestellt. Anschließend haben wir erklärt, dass Privatheitsbedrohungen nicht alleinig von den Tags und den preisgegebenen Orten ausgehen, sondern auch von den durch die mobile Tagging-Anwendung aufgezeichneten Tracks.

Trackaufzeichnung Zu Beginn dieser Phase haben wir einen Fragebogen ausgegeben, der nach allen Situationen und Orten fragt, an denen unsere Teilnehmer ihre Position nicht für andere sichtbar machen möchten. Anschließend haben wir die Privatheitsmechanismen aktiviert, die es den Teilnehmern erlauben, ihre Tracks zu verbergen. Wir haben die Teilnehmer auf den PETs trainiert, ihnen jedoch ganz bewusst nicht detailliert jede mögliche Bedrohung beschrieben. Da wie beschrieben einige der PETs nicht orthogonal sind, wie die Definition privater Bereiche (PET_{areas}) und der GPS-Schalter (PET_{switch}), haben wir entschieden, jedes PET isoliert für sich zu untersuchen. Um die korrekte Benutzung zu garantieren, haben wie mit dem PET begonnen, das die meiste Konfiguration benötigt. Diese Konfiguration haben wir dann während unseres Treffens mit den Teilnehmern gemeinsam vorgenommen. Das letzte getestete PET ist das am schwierigsten zu verstehende, jedoch das am einfachsten zu konfigurierende gewesen.

Wir haben mit PET_{areas} begonnen. Wir haben die Teilnehmer aufgefordert Bereiche zu definieren, in denen sie privat sein möchten, das heißt ihr Track nicht preisgegeben werden darf. Um zu vergleichen, ob die Teilnehmer die Bereiche, die ihren Präferenzen entsprechen, korrekt benennen können, haben wir anschließend jedem Teilnehmer seinen individuellen Track eingeblendet. Diesen haben wir vorab während der Tagging-

Phase im Hintergrund aufgezeichnet. Die Teilnehmer haben nun ihre Bereiche bei Bedarf anpassen können. In den folgenden zwei Tagen haben die Teilnehmer PET_{areas} genutzt, um in ihrem Alltag Positionen zu schützen, die sie als privat empfinden. Für weitere zwei Tage haben die Teilnehmer den GPS-Schalter nutzen können, der die Aufzeichnung des Tracks startet beziehungsweise unterbricht (PET_{switch}). Für die letzten beiden Tage haben wir die Teilnehmer den Anonymisierer (PET_{anon}) nutzen lassen.

Am Ende sowohl der Tagging- als auch der Trackaufzeichnungsphase haben die Teilnehmer die feingranulare Konfiguration (PET_{fine}) nutzen können. Mit PET_{fine} haben die Nutzer definieren können, wer welche aufgezeichneten Orte (Abschnitte von Tracks) sehen darf, die nicht bereits von einem der anderen Verfahren geschützt sind.

Abschluss Wir haben unsere Studie mit einem Fragebogen abgeschlossen. Gegenstand der abschließenden Befragung ist die Nutzung und Wahrnehmung der angebotenen Mechanismen gewesen.

4.2.4. Evaluation

In diesem Abschnitt beantworten wir die Forschungsfragen F1 – F4. Wir präsentieren unsere Ergebnisse zweigeteilt, entsprechend der Unterteilung der Studie in die Tagging- sowie die Trackaufzeichnungsphase.

4.2.4.1. Positionen, Inhalte und Metadaten

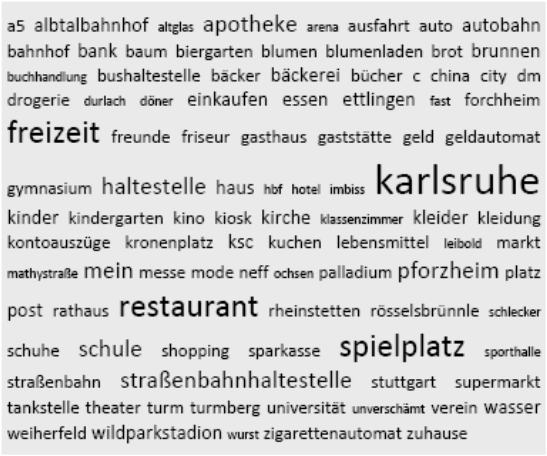
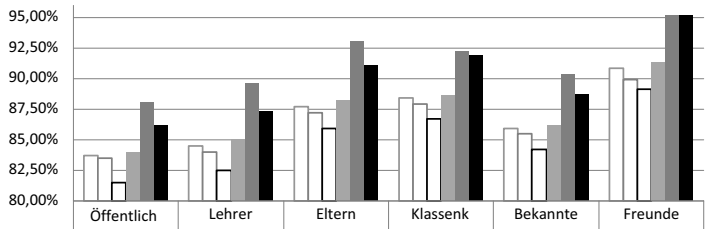


Abbildung 4.1.: Tag-Cloud annotierter Orte

Unsere Teilnehmer haben 1042 Tags an 442 Orten erstellt. Das heißt, sie haben 442 Tupel der Form `<position, tags, metadaten>` erstellt, die die Orte ihres alltägli-

chen Lebens explizit machen. Die meisten Orte sind mit zwischen einem und drei Tags annotiert worden. Zur besseren Einordnung dieses Verhaltens ist anzumerken, dass diese Beobachtung mit den Erfahrungen aus anderen Tagging-Anwendungen [MNBD06] übereinstimmt. Die Teilnehmer haben 41% der annotierten Orte mit zwei oder mehr Tags versehen. Eine Übersicht der Tags in Form einer Wolke (engl. tag cloud) gibt Abbildung 4.1.

Welche Informationen geben Personen im LBS Umfeld preis (F1)? Um herauszufinden, welche Informationen Personen gerne untereinander austauschen möchten, evaluieren wir, welche Tags und Metadaten unsere Teilnehmer mit Hilfe von $PET_{checkbox}$ und PET_{fine} öffentlich zugänglich gemacht haben. Unsere Hypothese lautet, dass Teilnehmer die meisten Informationen preisgeben möchten.



	Öffentlich	Lehrer	Eltern	Klassenk	Bekannte	Freunde
Datum	83,74%	84,48%	87,68%	88,42%	85,96%	90,89%
Zeit	83,50%	83,99%	87,19%	87,93%	85,47%	89,90%
Name	81,53%	82,51%	85,96%	86,70%	84,24%	89,16%
Pseudonym	83,99%	84,98%	88,18%	88,67%	86,21%	91,38%
Tags	88,08%	89,67%	93,05%	92,25%	90,37%	95,23%
Orte	86,13%	87,28%	91,04%	91,91%	88,73%	95,23%

Abbildung 4.2.: Preisgegebene Inhalte und Metadaten

Die Spalte ‘öffentlich’ in Abbildung 4.2 zeigt, wie oft Informationen der unterschiedlichen Kategorien preisgegeben worden sind. Die Abbildung unterscheidet zwischen Datum und Uhrzeit, zu der die Tags erstellt worden sind, dem Name und dem Pseudonym des Erstellers, dem Tag selbst und der Position des Tags. Die Abbildung zeigt, dass unsere Teilnehmer 88% aller generierten Tags, und 86% aller getaggten Positionen für jeden frei zugänglich machen wollen. 81% aller `<position, tags, metadaten>` Tupel sind als ’öffentlich’ markiert, nur 2% der Tupel haben die Teilnehmer vollständig privat gemacht.

Die Teilnehmer haben feingranulare Unterscheidungen bezüglich der preiszugebenden Information vorgenommen. Das bedeutet, dass sie nur einzelne Komponenten des Tupels preisgeben möchten. Für 10% aller Tupel sind die gleichzeitig preisgegebenen Tags zu einer Position nur anteilig ‘öffentlich’ gemacht worden. Zum Beispiel ist nur ’Schloss’ von ’Treffpunkt Freunde Schloss’ veröffentlicht worden. In 4% der Fälle haben die Teilnehmer nur die Position und die Tags preisgegeben, die Metadaten hingegen

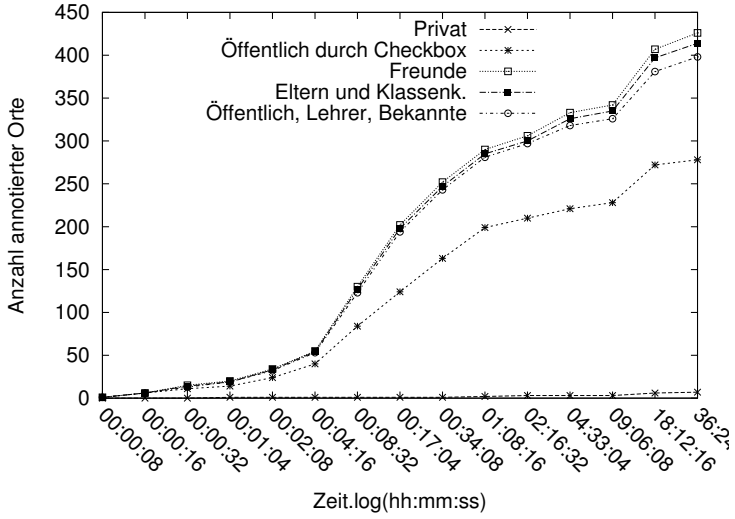


Abbildung 4.3.: Lebensmittelpunkte nach Zeitmessung – time.log(hh:mm:ss)

vollständig für sich behalten. Für 3% der Tupel teilen die Teilnehmer die Metadaten nur eingeschränkt mit anderen Personen, zum Beispiel das Pseudonym aber nicht die Zeit der Erstellung. Datum und Zeit der Erstellung und das Pseudonym des Erstellers sind nahezu gleich oft als 'öffentlich' markiert worden wie Tags. Den echten Namen haben die Teilnehmer hingegen am seltensten preisgegeben. Keiner unserer Teilnehmer hat berücksichtigt, dass die Verknüpfung mehrerer verschiedener Tags und Metadaten mit unterschiedlichen Privatheitseinstellungen unter Umständen Rückschlüsse auf Daten erlaubt, die sie eigentlich als privat definiert haben. Dies ist beispielsweise der Fall, wenn ein Teilnehmer auch nur ein einziges Mal sein Pseudonym zusammen mit seinem echten Namen preisgibt. Für alle anderen Positionen, die die Teilnehmer gezielt nur mit dem Pseudonym veröffentlicht haben, kann der Name so leicht zugeordnet werden.

Die versehentliche Preisgabe von Informationen an wichtigen Orten des alltäglichen Lebens kann eine wesentliche Bedrohung für die Privatheit darstellen. Wir wollen herausfinden, ob die Teilnehmer andere Personen die Zentren ihres Lebens einsehen lassen. Dazu haben wir die Zeit gemessen, die sich eine Person in der Nähe von von ihr getaggter und veröffentlichter Orte aufhält. Abbildung 4.3 zeigt die kumulative Verteilungsfunktion der Zeit in logarithmischer Darstellung. Wir haben die Zeit gemessen, wenn sich Teilnehmer näher als 200 Meter an von ihnen preisgegebenen Position aufhalten. Wie man ablesen kann, verweisen 20% aller Tupel auf Orte, an denen der Ersteller des Tags während des 14-tägigen Experiments mehr als 20 Stunden verbracht hat. Solche Orte umfassen das Schulgebäude oder das Zuhause der Teilnehmer. Man erinnere sich, dass unsere Teilnehmer 2% aller Tupel vollständig privat gemacht haben.

50% dieser Tupel beziehen sich auf Orte, an denen der Ersteller des Tags mindestens 20 Stunden verbracht hat.

Zwischenfazit: Obige Erkenntnisse führen zu der folgenden Schlussfolgerung: (1) Anstelle der Preisgabe eines Minimums an Information tendieren Personen dazu, alles an jeden preiszugeben. Sehr sensible Information wird hingegen sehr wohl privat gemacht. (2) Viele Nutzer bedenken nicht, dass die Verknüpfung von Informationen zu Privatheitsproblemen führen kann. Dies erfordert automatisierte PETs, die den Nutzer warnen, wenn er unbewusst Daten preisgibt, die zu einem umfangreichen Profil verknüpft werden können. (3) Auch wenn Personen viele Tupel gemäß 'Alles oder Nichts' preisgeben, so gibt es einen Bedarf an PETs, die feingranular einstellbar machen, welche Informationen anderen Personen zugänglich sein sollen. (4) Die Zeit, die die Teilnehmer im Umfeld eines preisgegebenen Tags verbracht haben, zeigt, dass Nutzer häufig wichtige Orte ihres täglichen Umfeldes preisgeben.

Welche sozialen Gruppen dürfen welche private Daten einsehen (F2)? PET_{fine} erlaubt es Nutzern, nicht nur die Art der preisgegebenen Daten zu unterscheiden, sondern auch zu bestimmen, welche Personengruppe die Daten sehen darf. Mögliche Gruppen im Experiment sind Lehrer, Eltern, Klassenkameraden, Bekannte und Freunde. Abbildung 4.2 zeigt, welche Personengruppen häufig welche Informationen sehen dürfen. Man erinnere sich, dass die Teilnehmer 81% aller Daten vollständig öffentlich gemacht haben, und 2% vollständig privat. Folglich hat jede Personengruppe wenigstens 81% aller Tupel sehen dürfen ('öffentlich'). Unsere Teilnehmer haben die meiste Information der Personengruppe Freunde zugänglich gemacht und am wenigsten der Gruppe Lehrer. Vergleichen wir die Aussagen der Teilnehmer im Fragebogen mit Abbildung 4.2 sehen wir ein paar Abweichungen: Im Fragebogen haben 6 der 10 Teilnehmer angegeben, keine Daten über ihren Aufenthaltsort an die Eltern weitergeben zu wollen, 2 Personen sind unentschieden gewesen, 2 geneigt, ihre Position preiszugeben. Unsere Ergebnisse aus der Studie haben jedoch gezeigt, dass die Eltern zwischen 86% und 93% aller Daten sehen dürfen, abhängig von der Art der Information. Für alle Personengruppen ist der Korrelationskoeffizient zwischen den preisgegebenen Metadaten und den preisgegebenen Orten 0.99, für Tags beträgt er 0.97.

Wir haben den Teilnehmern eine Liste von Beispielorten gezeigt, die sie taggen könnten. Diese umfasst unter anderem ihr Haus und das Haus von Freunden. Wie zu erwarten, haben alle Teilnehmer ihr Haus markiert, vier von ihnen haben das Tag sogar öffentlich gemacht. 8 von 10 Teilnehmern haben mindestens ein Haus eines Freundes getaggt. Jedoch nur ein Teilnehmer hat dieses Tag öffentlich zugänglich gemacht, einer der Gruppe Freunde. Das stimmt mit den Ergebnissen aus unserem vorangegangenen Fragebogen überein, in dem 8 der Teilnehmer angegeben haben, sich Gedanken über die Privatheit von Freunden zu machen.

Abbildung 4.3 gibt die Zeit an, die die Teilnehmer nahe ihrer getaggten Orte verbracht haben. Außerdem unterscheiden wir, welche Personengruppen diese Tags haben sehen dürfen. Folglich gibt die Abbildung an, welche Personengruppen die Aufenthaltsorte der Tag-Ersteller beobachten können, darunter auch die Lebensmittelpunkte der Teilnehmer. Zur übersichtlicheren Darstellung haben wir Kurven mit sehr ähnlichen Werten zusammenfasst. 16% aller `<location, tags, metadata>`-Tupel geben Auskunft darüber, wo der Ersteller des Tags sich mehr als 20 Stunden aufgehalten hat. Die Freunde des jeweiligen Teilnehmers dürfen 95% aller Orte sehen, davon sind 19% Orte, an denen der Ersteller des Tags mehr als 20 Stunden verweilt hat.

Zwischenfazit: Zusammengefasst (1) unterscheiden Personen zwischen unterschiedlichen Personengruppen (hier für 17% der generierten Daten). Bezogen auf die Preisgabe der Metadaten zu den unterschiedlichen sozialen Gruppen (2) verhalten sich Nutzer sehr ähnlich. Darüber hinaus (3) sorgen sie sich um die Privatheit bestimmter Personengruppen mehr als über ihre eigene.

Wie nutzen Personen PETs und welche PETs bevorzugen sie (F3, F4)? Unsere Teilnehmer haben drei Optionen gehabt, ihre Privatheit zu schützen: (i) sie haben die Checkbox auf dem mobilen Endgerät nutzen können, um ein vollständiges Tupel privat / öffentlich zu machen ($PET_{checkbox}$). (ii) Sie haben unsere Webseite nutzen können, um zu bestimmen, wer genau welche Information sehen darf (PET_{fine}). (iii) Die Teilnehmer haben darauf verzichten können, Informationen preiszugeben, wenn sie es für zu gefährlich halten. Wir haben beobachtet, dass die Teilnehmer 63% aller Tupel mit $PET_{checkbox}$ öffentlich sichtbar gemacht haben. Für 37% der Tupel wurden die Privatheitseinstellungen mit Hilfe von PET_{fine} definiert. Die Teilnehmer haben PET_{fine} genutzt, um weitere 23% aller Positionsinformationen öffentlich zu machen.

Zwischenfazit: Wir folgern daraus, dass Nutzer zwar einfache Mechanismen bevorzugen, sehr wohl aber auch ausgefeilte PETs einsetzen. Darüber hinaus sind Personen bereit, einen relativ hohen Aufwand zu betreiben, um sensible Informationen, zum Beispiel das eigene Haus oder das von Freunden, zu schützen. Aus diesem Grund sehen wir einen klaren Bedarf an PETs, die detaillierte Privatheitseinstellungen erlauben.

4.2.4.2. Tracks

Dieser Abschnitt beschreibt den Vergleich der Nutzung von PET_{areas}, PET_{switch} und PET_{anon}, gefolgt von einer Analyse der Tracks, die die Teilnehmer öffentlich gemacht haben.

Wie werden PETs genutzt und was geben Nutzer preis (F1, F4)? Um herauszufinden, wie die Teilnehmer PETs einsetzen, betrachten wir die aufgezeichneten Tracks.

Tabelle 4.5.: PET_{areas}: Private Bereiche initial und überarbeitet

Initial	Jeder	Lehrer	Eltern	Freunde	Verwandte	Klassenk.
Anzahl Bereiche	12	23	16	16	18	19
Avg. Größe (km^2)	0.712	1.121	0.578	1.576	0.504	1.339
Stddv. Größe	1.191	3.496	1.051	4.146	1.007	3.826
Zeit in Bereich	24.5%	26.9%	24.8%	26.9%	25.0%	26.5%
Überarbeitet	Jeder	Lehrer	Eltern	Freunde	Verwandte	Klassenk.
Anzahl Bereiche	18	28	23	20	22	25
Avg. Größe (km^2)	0.984	0.720	0.823	0.966	0.832	0.806
Stddv. Größe	1.494	1.258	1.356	1.418	1.387	1.307
Zeit in Bereich	33.2%	40.2%	33.9%	34.7%	34.0%	34.3%

PET_{areas} (Private Bereiche) Dieses PET erlaubt es Personen, Bereiche zu definieren, in denen andere Personen sie nicht sehen können, beispielsweise 'überhaupt niemand'. Wir werden die Bereiche untersuchen, die wir jeweils am Anfang und am Ende der Trackaufzeichnungsphase protokolliert haben.

Zu Beginn haben wir einen Fragebogen ausgegeben, mit dessen Hilfe die Teilnehmer ihre privaten Bereiche haben definieren können. Die Teilnehmer mussten zuerst beschreiben, in welchen Situationen sie nicht gesehen werden wollen und dann, welche Orte das umfasst. Anschließend haben sie ihre jeweiligen zu schützenden Bereiche über ein von uns implementiertes Google-Mashup auf eine elektronische Karte abgetragen. Unsere Teilnehmer haben 26 private Bereiche definiert. Diese schützen Freizeitaktivitäten (11), das Zuhause (6), die Schule (5), den (die) Freund(in) (2), die Arbeit(1) und Verwandte (1). Tabelle 4.5 zeigt, wie viele Bereiche definiert wurden, deren durchschnittliche Größe (Avg.) und Standardabweichung (Stddv.) sowie die Zeit, die die Teilnehmer innerhalb der geschützten Bereiche verbracht haben. Um unsere Statistik nicht zu verfälschen, haben wir zwei Bereiche entfernt, die einmal das ganze Land und einmal die ganze Welt umspannen. Diese Bereiche wurden auch von den Teilnehmern bei der Nutzung des Dienstes entfernt. Um zu überprüfen, ob die Bereiche die Lebensmittelpunkte der Teilnehmer umfassen, haben wir außerdem die Zeit gemessen, die sich die Teilnehmer in mindestens einem ihrer Bereiche aufgehalten haben.

Die Bereiche, die die Teilnehmer initial definiert haben, decken die Aufenthaltsorte der Teilnehmer so ab, dass sie im Mittel 25% der Experimentzeit vor jedem geschützt sind. Nachdem wir den Teilnehmern ihre Tracks aus der Tagging-Phase eingeblendet und die Teilnehmer außerdem unsere Anwendung zwei weitere Tage im Einsatz hatten, haben wir die von den Teilnehmern angepassten Bereiche protokolliert. Die Teilnehmer haben 6 neue Bereiche angelegt, die eine Preisgabe gegenüber jedem verhindern. Im Mittel haben unsere Teilnehmer außerdem die Größe ihrer Bereiche verdoppelt. Ein

Tabelle 4.6.: Vergleich von PET_{switch} und PET_{areas}

GPS / Position	in privatem Bereich	außerhalb eines Bereichs	Summe
Aktiv (%)	43.56%	31.71%	75.27%
Inaktiv (%)	16.45%	8.27 %	24.73%
Summe (%)	60.01%	39.99%	**100.00**%

Bereich ist um Faktor 16 vergrößert worden, nachdem der Teilnehmer bemerkt hat, dass es ein Bereich exakt über dem eigenen Haus anderen Personen erlaubt zu erkennen, wann er das Haus betritt und wann er es wieder verlässt. Alles in allem haben die Teilnehmer 29 Bereiche definiert, die sie auch hin und wieder betreten haben. Der Vergleich der Zeit, die die Teilnehmer in den überarbeiteten Bereichen verbracht haben mit der Zeit in den ursprünglichen Bereichen (und gemessen über den gesamten Experimentzeitraum) zeigt, dass die Teilnehmer nun 33% der Experimentzeit schützen. Vorher sind das nur 25% gewesen.

Zwischenfazit: Unsere Ergebnisse zeigen, dass Nutzer ihre Privatheitseinstellungen häufig anpassen möchten. Außerdem können sich LBS-Nutzer Bedrohungen nicht vorstellen. Die Tatsache, dass die Teilnehmer nach dem Einblenden ihrer individuellen Tracks grundlegende Änderungen an ihren Bereichen vorgenommen haben, unterstützt diese Aussage. Das heißt, im Kontext von LBS bedarf es der (zum Beispiel) visuellen Aufbereitung der Daten, bevor PETs richtig eingesetzt werden können.

PET_{switch} (GPS-Schalter) Der zweiter Mechanismus, den wir evaluiert haben, ist der manuelle Schalter zum Aktivieren / Deaktivieren des GPS-Empfängers. Wir haben mitgeschrieben, wann und wo das GPS ausgeschaltet wurde, also wann und wo die Teilnehmer ihre Position nicht haben preisgeben wollen. Um PET_{switch} zu evaluieren nehmen wir an, dass die Bereiche, die die Teilnehmer zuvor im Fragebogen definiert und dann in unser System übertragen und auch verwendet haben, eine gute Beschreibung der Privatheitspräferenz des jeweiligen Teilnehmers darstellen.

Der GPS-Schalter ist jedoch nur selten genutzt worden. Während der zwei Tage haben die Teilnehmer den Schalter im Mittel nur jeweils 6-mal benutzt. Jeder von ihnen hat den Schalter weniger als 12-mal benutzt. Wie Tabelle 4.6 zeigt, haben die Teilnehmer im Mittel 44% der Tracks sichtbar gemacht, während sie sich in einem geschützten Bereich aufgehalten haben und eigentlich unbeobachtet sein wollten. Das zeigt deutlich, dass PET_{switch} in der Praxis versagt.

Zwischenfazit: PETs, die eine andauernde Aufmerksamkeit des Anwenders erfordern, führen häufig zu einer ungewollten Preisgabe von Informationen. Das gilt beispielsweise, wie hier untersucht, für einfache Ein- / Ausschalter für das GPS. Wichtig zu

Tabelle 4.7.: PET_{anon}: Mittlere Distanz zu den k-nächsten Nachbarn

k	2	3	4	5	6
Distanz (km)	1.06	1.67	2.00	2.35	2.72
k	7	8	9	10	–
Distanz (km)	3.17	3.83	5.79	13.43	–

erwähnen ist, dass die Teilnehmer diese Erkenntnis auch in unserem abschließenden Fragebogen bestätigt haben.

PET_{anon} (Anonymisierer) Dieses PET adaptiert k-Anonymität, um zu verhindern, dass Sequenzen von Positionen zu einem Track zusammengeführt werden können. Wichtig bei diesem PET ist, dass die Wahl von k weitgreifende technische Kenntnisse erfordert. Aus diesem Grund haben wir PET_{anon} im Detail erklärt. Man erinnere sich, dass die Teilnehmer davon ausgegangen sind, dass auch andere Personen diese Anwendung nutzen. Deshalb haben die Teilnehmer auch Werte für $k > 10$ einstellen können.

Unsere Teilnehmer haben ein k gewählt, das zwischen 1, das heißt keine Anonymität, und 30 variiert. Das mittlere k ist 7,33 gewesen, mit einer Standardabweichung von 8,64. Um herauszufinden, welches k angemessen ist, haben wir den gegenseitigen Abstand der 14.015 aufgezeichneten Positionsinformationen berechnet. Als ein Beispiel zeigt Tabelle 4.7, dass die mittlere Distanz zum drittnächsten Nachbarn 1,67km ist. Folglich bedeutet eine Position, die mit k=3 anonymisiert wird, dass ein Nutzer im Mittel in einem Radius von $8,87km^2$ nicht erkannt wird[7]. Das von den Teilnehmern gewählte mittlere k entspricht einem Kreis von $38,5km^2$, was deutlich über der Größe der Bereiche liegt, die die Teilnehmer mit PET_{areas} definiert haben.

Zwischenfazit: Wir schließen daraus, dass Nutzer komplexe Techniken, wie den Anonymisierer, nicht in Übereinstimmung mit ihren Präferenzen nutzen können.

Welche Personengruppen dürfen die Tracks sehen (F2)? PET_{fine} und PET_{areas} erlauben die teilweise Einschränkung der Sichtbarkeit von Tracks für unterschiedliche Personengruppen. Am Ende der Studie haben 5 Teilnehmer PET_{fine} eingesetzt, um für 270 Trackabschnitte spezielle Privatheitspräferenzen zu spezifizieren. Mit der Preisgabe dieser Tracks wollten sie zum Beispiel den Besuch einer Diskothek kundtun. Die andere Hälfte der Teilnehmer hat keine weitere Track-Information preisgegeben.

Die Teilnehmer haben nur 28% der Tracks für andere sichtbar gemacht. Tabelle 4.8 zeigt, an welche sozialen Gruppen die Teilnehmer diese 28% preisgegeben haben. Sie

[7] Wir haben kreisförmige Bereiche angenommen, um die Funktionalität leichter zu vermitteln. [MCA06] nutzt ein Gitter (engl. Grid) aus unterschiedlich großen rechteckigen Zellen, die zu einer Pyramide angeordnet sind.

Tabelle 4.8.: Preisgabe von Tracks über PET_{fine}

Phase	öffentlich	Lehrer	Eltern	Freunde	Bekannte	Klassenk.
Tagging (%)	2%	4%	13%	28%	7%	10%
Track (%)	0%	0%	8%	23%	2%	4%
Summe (%)	2%	4%	21%	51%	9%	14%

haben nur 2% aller Tracks öffentlich, also jedem zugänglich gemacht. 4% der Tracks dürfen Lehrer, 21% Eltern und 51% Freunde sehen.

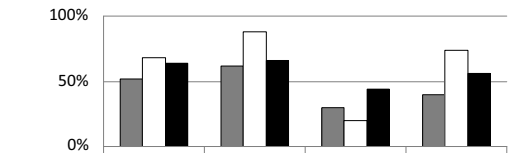
PET_{areas} ermöglicht es Nutzern zu konfigurieren, wer die Daten, die innerhalb eines privaten Bereiches erhoben worden sind, nicht sehen darf. Tabelle 4.5 vergleicht für jede der sozialen Gruppen die Anzahl der definierten Bereiche, ihre Größe und die Zeit, die die Teilnehmer darin verbracht haben.

Die Teilnehmer haben die meisten Bereiche definiert, um Lehrern keinen Einblick in ihre Bewegungsprofile zu geben, und am wenigsten Bereiche, um sich vor Freunden zu schützen. Das stimmt auch mit den Ergebnissen von PET_{fine} überein. Betrachtet man die Auswirkungen der Bereiche zeitlich, so schützen sich die Teilnehmer für 40% der Experimentzeit vor Lehrern. Für alle anderen Gruppen schützen sie sich für weniger als 35% der Zeit.

Zwischenfazit: Unsere Evaluierung führt zu zwei Erkenntnissen: Erstens scheinen Personen besorgter um die Preisgabe von Tracks als um die Preisgabe von einzelnen Positionen, Inhalten (Tags) und Metadaten zu sein. Während die Teilnehmer 81% von Letzterem öffentlich machen, so geben sie fast keine Trackabschnitte an jeden beliebigen weiter. Zweitens gibt es einen Unterschied zwischen der Track-Information, die mit PET_{fine} und PET_{areas} preisgegeben worden ist. Diese Erkenntnis stimmt auch mit unseren gewonnenen Erfahrungen aus der Evaluierung von PET_{switch} überein.

Welche Art PET bevorzugen die Nutzer (F4)? Um diese Frage zu beantworten, haben wir den finalen Fragebogen ausgewertet. Wir haben die Teilnehmer aufgefordert, eine Bewertung der PETs auf einer 5-Punkte-Likert-Skala für die Kriterien Sicherheit, Leichtigkeit der Nutzung, Komplexität des Verständnisses und Aufwand der Nutzung für die PETs PET_{areas}, PET_{switch} und PET_{anon} vorzunehmen. Die Mittelwerte aus den Bewertungen sind in Abbildung 4.4 dargestellt. Außerdem haben wir die Teilnehmer aufgefordert PETs oder PET-Kombinationen zu bewerten, die sie in Zukunft nutzen wollen. Tabelle 4.9 zeigt den Mittelwert der Bewertung.

Zwischenfazit: Unter allen PETs isoliert betrachtet, haben unsere Teilnehmer PET_{switch} bevorzugt. Sie haben diesem PET in allen Kategorien die beste Bewertung gegeben. PET_{areas} hat in allen Kategorien die schlechteste Bewertung erhalten. Die Teilnehmer



	Sicherheit	Einfache Nutzung	Schwierig zu verstehen	Geringer Aufwand
PETareas	52%	62%	30%	40%
PETswitch	68%	88%	20%	74%
PETanon	64%	66%	44%	56%

Abbildung 4.4.: Vergleich der Mechanismen

Tabelle 4.9.: PET-Bewertungen

PET Bewertung	∅
GPS Schalter & Anonymisierer	2,9
Alle kombiniert	3,1
Private Bereiche & GPS Schalter	3,4
Private Bereiche & Anonymisierer	3,9
GPS Schalter	4,2
Anonymisierer	4,8
Private Bereiche	5,4

haben jedoch wahrgenommen, dass der GPS Schalter kontinuierlich Aufmerksamkeit erfordert. Aus diesem Grund wünschen sie eine Kombination mit anderen PETs. In unserem Fall ist das die Kombination von PET_{switch} und PET_{anon}. Wir finden das überraschend, da die Teilnehmer weder den GPS Schalter als auch den Anonymisierer gemäß ihrer Präferenzen eingesetzt haben. Eine mögliche Erklärung ist, dass die Teilnehmer das einfachste und intuitivste PET mit dem PET kombinieren wollten, dass automatisiert arbeitet und nur die Definition eines Parameters erfordert.

4.2.4.3. Diskussion

Es ist unsere Entscheidung gewesen, eine sorgfältig geplante Studie, unter echten Bedingungen und mit einer begrenzten Anzahl gut vorbereiteter Teilnehmer durchzuführen. Wir halten unsere Studienergebnisse aus drei Gründen für repräsentativ:

Relevanz der Teilnehmergruppe Unsere Studie umfasst 25 Personen, ihre Eltern und Lehrer. Wir haben 10 von ihnen mit mobilen Geräten ausgerüstet. Unsere Teilnehmer sind an neuen Technologien interessiert und alle sind den Umgang mit Mobiltelefonen und Internetanwendungen gewöhnt. Außerdem sind sie mobil, das heißt sie sind viel, jedoch in einem einigermaßen überschaubaren Umfeld unterwegs. Ein Gegenbeispiel wäre ein Büroarbeiter, der 10 Stunden am Tag ohne Kontakt zu einer anderen sozialen Gruppe als 'Kollegen' seinen Standort

nicht ändert. Des weiteren stellen die Teilnehmer eine relevante Zielgruppe für Mobilfunkanbieter dar.

Voll funktionsfähige, reale Anwendung Wir haben unsere Erkenntnisse mit einer Geo-Tagging-Anwendung erzielt. Unsere Anwendung ist ähnlich aller LBS, bei denen Nutzer standortbezogene Inhalte zu Positionen, an denen sie vorher gewesen sind, untereinander austauschen. Die Teilnehmer haben Verhaltensmuster gezeigt, die denen aus [MNBD06] ähnlich sind. Am Ende unserer Studie haben einige Teilnehmer gefragt, ob sie unsere Anwendung weiterhin nutzen könnten.

Echte Bedrohungen der Privatheit Die Teilnehmer haben unsere Anwendung in ihren Alltag integriert. Die Studie hat Arbeitstage wie auch Feiertage umfasst. Die mittlere Nutzungsdauer unserer Anwendung ist 6 Stunden am Tag gewesen, das ist unserer Ansicht nach sehr hoch. Während der Studie haben die XDAs GPS Daten übermittelt, die deutlich besser als auf 100 Meter genau gewesen sind. Die Auflösung ist so gut, dass wir jedes einzelne betretene Gebäude identifizieren können.

Die drei markantesten Erkenntnisse aus der Durchführung der Studie sind:

Personengruppen Die Teilnehmer wollten die generierten Informationen austauschen, vielfach jedoch nur mit einem begrenzen Personenkreis, zum Beispiel Eltern oder Freunden. Wir empfehlen aus diesem Grund PETs, die eine Unterscheidung zwischen sozialen Gruppen erlauben. Was in sozialen Netzwerkseiten mittlerweile üblich ist, stellt bei Web 2.0-Diensten, die kontinuierlich und im Verborgenen arbeiten, hohe Anforderungen an die Nutzer und die Entwickler von PETs.

Unterschiedliche Informationen LBS verarbeiten nicht nur Positionen und Tracks, sondern auch andere Arten von Information wie Inhalte und Metadaten. PETs für LBS müssen solche Informationen berücksichtigen, da Privatheitsbedrohungen von jeder Art dieser Daten ausgehen können. Insbesondere gilt dies auch für die Kombination unterschiedlicher Daten. Hinzu kommt, dass Nutzer häufig nur Teile der Informationen preisgeben.

Komplexe PETs Aus Gründen der Transparenz wird häufig angenommen, dass Nutzer leicht verständliche PETs bevorzugen. Wir haben jedoch gezeigt, dass die Teilnehmer komplexe PETs einfachen vorziehen, wenn das einfache PET zur Durchsetzung der Privatheitspräferenzen eine kontinuierliche Aufmerksamkeit und Aufwand erfordert.

4.2.5. Zusammenfassung

Standortbezogene Dienste (LBS) sind eine wichtige aktuelle Entwicklung. Jedoch setzen LBS die Privatheit ihrer Nutzer einem hohen Risiko aus. Wir haben untersucht, welche Privatheitsmechanismen Personen nutzen und insbesondere, wie sie dies tun. Zu diesem Zweck haben wir eine voll funktionsfähige Geo-Tagging-Anwendung ent-

wickelt und darauf aufbauend eine Nutzerstudie durchgeführt. Unsere Studienteilnehmer haben diesen LBS in ihren Alltag integriert.

Neben anderen Ergebnissen haben wir herausgefunden, dass Personen dazu tendieren, Mechanismen einzusetzen, die leicht zu verstehen sind. Erfordern diese Mechanismen jedoch eine kontinuierliche Aufmerksamkeit von dem Nutzer, versagen sie in der Praxis. Das heißt, dass ein Nutzer mit diesen Mechanismen seine Privatheitspräferenzen nicht durchsetzen kann. Aktuell setzt Google Latitude solche ein PET ein.

Aus diesem Grund wollen Nutzer einfache Mechanismen mit solchen kombinieren, die automatisiert arbeiten und eine feingranulare Entscheidung erlauben, wer welche Informationen sehen darf. Obwohl Nutzer dazu tendieren, die meiste Information (hier 81%) vollständig preiszugeben, so haben sie doch für einige Daten (17%) auf sehr detaillierte Art und Weise spezifiziert, wer diese Informationen sehen darf. Dies beinhaltet insbesondere auch sensible Informationen wie den Ort der eigenen Wohnung. Schließlich haben wir herausgefunden, dass Personen mit der Privatheit bestimmter Personengruppen vorsichtig umgehen, auch wenn sie über sich selbst alles preisgeben.

4.3. Fazit

Wir haben zwei Web 2.0-Dienste, repräsentativ für viele relevante, gerade erst den Markt durchdringende Anwendungen untersucht. Wir haben beide Dienste von Grund auf entwickelt. Mittels darauf aufbauenden Nutzerstudien haben wir untersucht, welchen Anforderungen ein PET genügen muss, damit es die Privatheit von Nutzern untereinander erfolgreich schützt. Im Folgenden fassen wir die Ergebnisse, auch aus Kapitel 3 zusammen. Wir unterscheiden dabei inhaltlichen Anforderungen, Anforderungen an die Benutzbarkeit und die funktionalen Anforderungen:

Inhaltliche Anforderungen Inhaltliche Anforderungen beziehen sich auf 'was', 'wann', 'vor wem' geschützt werden soll, also den Inhalt der Privatheitspräferenzen. Nutzer geben die Mehrheit von Informationen preis, machen bei sensiblen Daten jedoch feingranulare Unterscheidungen. Ein PET muss zwischen unterschiedlichen Personengruppen unterscheiden können. Weiter muss ein PET die Möglichkeit bieten, für unstrukturiert (vgl. Definition 3 Kapitel 3) preisgegebene Inhalte, hier zum Beispiel Suchanfragen oder Tags, Privatheitspräferenzen zu definieren. Weiter muss ein PET unterschiedliche Situationen erkennen können, beispielsweise 'beim Arbeiten', und zeitliche Einschränkungen der Sichtbarkeit von Daten erlauben. Außerdem muss ein PET die Formulierung reziproker Strategien ermöglichen. [10] untersucht den Aspekt der Reziprozität im Kontext von Vertrauensstrategien genauer. Darüber hinaus muss ein PET die Anonymisierung der Daten ermöglichen.

Anforderungen an die Benutzbarkeit Ein PET muss den Aufwand für den Nutzer minimieren. Nutzer wollen einfache PETs, können mit diesen jedoch nicht umge-

hen, wenn die PETs eine kontinuierliche Aufmerksamkeit von ihnen erfordern. Ein PET muss vom Nutzer verstanden werden, damit er es korrekt einsetzen kann. Wirkt ein PET jedoch automatisiert, vertrauen Nutzer auch komplexen, schwer verständlichen PETs. Eine Lösung könnte die Kombination komplexer, automatisiert arbeitender PETs sein mit einfachen PETs, über die Nutzer beispielsweise pauschal die Datenübermittlung deaktivieren können. Außerdem fordern Nutzer, dass sie Strategien sowohl positiv definieren können, also was sie jemandem erlauben, als auch negativ, das heißt, was sie jemandem verbieten.

Funktionale Anforderungen Die funktionalen Anforderungen beziehen sich auf die konkrete Ausgestaltung des PETs, also das 'wie'. Ein PET soll preisgegebene Daten lokal protokollieren und den Nutzer bei einer ungewollten Preisgabe informieren. Es soll im Hintergrund arbeiten und ohne den Dienstgeber funktionieren. Es muss mindestens Strategien der Form [ALWAYS | IF <*conditions*>] [DO NOT] DISCLOSE <*objects*> [TO <*groups*>] auswerten können. Wo möglich, sollte es eine visuelle Darstellung der Daten anbieten, beispielsweise preisgegebene Tracks auf Karten darstellen. Es muss die Herausforderung aufgreifen, die in den inhaltlichen Anforderungen geforderten Personengruppen zu definieren. Schwierig ist dabei die Unterscheidung zwischen Gruppen mit aufzählbaren Teilnehmern, wie Mitglieder der Familie, die ein Nutzer eindeutig benennen kann, und nicht aufzählbaren Gruppen, wie Kinder, Erwachsene oder Männer. Ebenso eine Herausforderung sind Präferenzen bezüglich des preisgegebenen Inhalts. Ein PET muss es einem Nutzer ermöglichen, Klassen von Inhalten zu schützen, beispielsweise über in Konzepte zusammengefasste Tags, Suchterme oder Links. Ein solches Konzept könnte 'Krankheit' sein. Offensichtlich ist es schwierig, alle konkreten Instanzen des Konzeptes 'Krankheit', wie 'Krebs', 'Bestrahlung' oder Medikamentenamen zu identifizieren und eindeutig einem Konzept zuzuordnen.

Mit den beiden Arbeiten haben wir eine Vielzahl wesentlicher Anforderungen an PETs identifiziert und mit Hilfe von Nutzerstudien evaluiert. Uns bekannte rudimentäre PETs bieten jeweils nur eine Teilmenge der geforderten Funktionalität an. Wir haben jedoch gesehen, dass all diese Anforderungen in Kombination auftreten. Wir haben uns in beiden Studien entschieden, gut geschulte Teilnehmer einzusetzen. Spannend für eine zukünftige Arbeit wäre es zu sehen, wie ein PET, das den oben genannten Anforderungen gerecht wird, von einer großen Gruppe ungeschulter Personen genutzt wird.

Weitere Erkenntnisse im Hinblick auf die folgenden Kapitel Neben den oben evaluierten Privatheitspräferenzen zwischen Nutzern haben wir die Nutzer in den Abschlussumfragen aufgefordert, ihre Privatheitspräferenzen gegenüber den Anbietern darzulegen. Wir haben die Teilnehmer auch hier nicht besonders auf Privatheitsaspekte trainiert, sondern sie eher nach ihrem Empfinden befragt.

Tabelle 4.10.: Privatheitsbedrohungen Nutzer–Suchmaschinenanbieter

Bedrohung	Häufigkeit
Werbung	40%
Profilbildung	23%
Weitergabe und Verkauf	23%
Spam	6%
Verfolgung illegaler Downloads	3%
Diskriminierung	3%

52% der Nutzer empfinden nach unseren Ergebnissen der CSE-Studie die Nutzung von Suchhistorien durch die Provider als problematisch (gar nicht problematisch 21%, eher unproblematisch 31%, eher problematisch 28%, sehr problematisch 21%). Die Probleme, die Nutzer dabei sehen, sind in Tabelle 4.10 zusammengefasst. Die am Häufigsten genannte negative Konsequenz ist (personalisierte) Werbung (40%), gefolgt von der Bildung von Persönlichkeitsprofilen (23%) sowie der Weitergabe und Verkauf der Daten (23%).

64% der Teilnehmer lesen vor der Anmeldung bei einem Dienst keine Datenschutzerklärung, 30% haben noch nich eine solche Erklärung gelesen. 73% der Teilnehmer, die eine Datenschutzerklärung gelesen haben, tun dies um herauszufinden, ob Daten weitergegeben werden. Weiter halten 70% der Teilnehmer, gemessen auf einer 5-Punkte-Likert-Skala, die Weitergabepraxis von Unternehmen für nicht nachvollziehbar. Auch haben 78% der Teilnehmer bereits eine Registrierung bei einem Dienst abgebrochen, weil dieser zu viele Daten vom Nutzer erheben wollte.

Die Teilnehmer der Studie zu standortbezogenen Diensten würden einen kommerziellen Anbieter nicht (80%) oder nur gegen eine hohe Ausgleichszahlung akzeptieren, beispielsweise 500 EUR für die Daten über einen Monat und 1200 EUR für ein Jahr. Das liegt deutlich über dem Wert unserer Frei-SMS.

Damit sich ein Dienst durchsetzen kann, müssen die Nutzer Vertrauen in den Anbieter haben. Wir werden in den folgenden zwei Kapiteln untersuchen, ob (i) das fehlende Vertrauen der Nutzer in die Anbieter berechtigt ist, (ii) aufzeigen, wie Nutzer Mängel der Datenschutzpraktiken von Anbietern effektiv identifizieren können und (iii) einen möglichen Ausweg für Nutzer wie Anbieter mittels Anonymisierung evaluieren.

5. Identifikation von Datenschutzverstößen der Diensteanbieter

Vertrauen ist die Grundlage aller Geschäftsbeziehungen [Xam09]. Datenschutz und Privatheit haben sich zu einem Reputationsfaktor für die Unternehmen entwickelt. Um die Kundenzufriedenheit zu steigern, haben viele Unternehmen Verhaltenskodizes für den Umgang mit personenbezogenen Daten entwickelt [WLW98]. Trotzdem fehlt es Nutzern an Vertrauen in die Anbieter von Web 2.0-Diensten. Das belegen die Ergebnisse der bisher beschrieben Studien [3, 4, 7]. Untermauert wird dies durch eine Vielzahl von Nachrichtenmeldungen über Datenschutzverstöße[1].

In diesem und dem folgenden Abschnitt konzentrieren wir uns auf den Schutz der Privatheit der Nutzer vor den Anbietern. Der vertrauensvolle Umgang der Anbieter mit den personenbezogenen Daten ist eine zentrale Voraussetzung, auch damit die bis hierhin untersuchten Mechanismen zum Schutz der Privatheit zwischen Nutzern überhaupt wirken können. Anders ausgedrückt, was hilft es, dass Nutzer die Sichtbarkeit ihrer Informationen für andere Nutzer einschränken können, wenn der Anbieter den gesamten Datenbestand preisgibt [PCT06] oder verkauft (siehe dazu auch den Disput zwischen Facebook und der Bundesministerin Ilse Aigner [Aig10]). Wir prüfen, ob das mangelnde Vertrauen der Nutzer in die Anbieter berechtigt ist. Das ist nach unserer Bewertung mindestens dann der Fall, wenn die Minimalanforderungen an die Datenschutzpraktiken der Unternehmen, das heißt die Anforderungen resultierend aus der Datenschutzgesetzgebung, nicht eingehalten werden.

Wir werden in Abschnitt 5.1 untersuchen, inwieweit sich Anbieter, insbesondere auch solche, die Web 2.0-Technologien einsetzen, konform zu geltendem Recht verhalten. Abschnitt 5.2 stellt einen Ansatz vor, wie alle am Austausch personenbezogener Daten beteiligten Parteien kollaborativ Fehlverhalten identifizieren können – und das ohne datenschutzrechtliches Expertenwissen.

[1] Breach Database, `http://www.idtheftcenter.org`, Mai 2010

5.1. Anbieterstudie zur Vollzugsdefizitanalyse

5.1.1. Ursache der Nutzer-Anbieter-Privatheitsprobleme

In diesem Abschnitt untersuchen wir, inwieweit der angesprochene Mangel an Vertrauen der Nutzer in die Anbieter berechtigt ist. Wir schließen die Ursachen, die aus fehlendem Bewusstsein eines Nutzers entstehen können, aus und analysieren, ob ein vollkommen bewusster Nutzer in der Lage ist, bei der Interaktion mit Web 2.0-Diensten die Kontrolle über seine personenbezogenen Daten zu wahren. Die zentrale Voraussetzung dafür ist die Einhaltung des Datenschutzrechts.

Eine mögliche Ursache für die Vielzahl der Datenschutzverstöße bei Online-Diensten sehen Experten in einem Vollzugsdefizit des Datenschutzrechts [WLW98]. Das heißt, Gesetze existieren, die Aufsichtsbehörden kontrollieren deren Einhaltung und Durchsetzung jedoch nicht ausreichend. Wir werden diese Hypothese in diesem Abschnitt evaluieren.

Um das Ausmaß des angenommenen Vollzugsdefizits zu untersuchen, haben wir in einer interdisziplinären Studie mit Juristen der Arbeitsgruppe von Prof. Kühling (Universität Regensburg) einen Katalog von möglichen Datenschutzverstößen bei Online-Diensten entwickelt. Anhand 100 ausgewählter Anbieter haben wir die von extern erkennbare Datenschutzpraktik der Anbieter mit dem Verstoßkatalog abgeglichen. So haben wir Verstöße identifiziert, aber auch aufgezeigt, welche Hürden es für den Nutzer gibt, einmal preisgegebene Daten aus dem Gedächtnis der digitalen Welt zu entfernen. Der Schwerpunkt innerhalb dieser Arbeit hat dabei auf der Methodik der Studie, der Auswahl der Anbieter, Verstöße beim Einsatz automatisierter Verfahren sowie bei dem Auskunfts- und Löschersuchen gelegen.

Unsere Ergebnisse sind alarmierend: Nur fünf der 100 untersuchten Anbieter verhalten sich konform zu geltendem Recht. Insgesamt haben wir bei 100 Anbietern mehr als 300 Verstöße identifiziert. Damit haben wir die Hypothese des Vollzugsdefizits belegt [6].

Wir stellen im Folgenden die angewandte Methodik der Studie vor (Abschnitt 5.1.2), gefolgt von unseren Ergebnissen (Abschnitt 5.1.3). Detailliertere Beschreibungen des rechtlichen Hintergrunds, beispielsweise zu Informationspflichten, Einwilligung und Widerruf, Auskunfts- und Löschersuchen etc. finden sich in Abschnitt 2.2. Wir werden hier nicht mehr näher darauf eingehen, nennen aber vor jedem evaluierten Aspekt die relevante rechtliche Grundlage.

5.1.2. Methodik der Anbieterstudie

In diesem Abschnitt beschreiben wir zunächst die Auswahl der Anbieter für unsere Studie (Abschnitt 5.1.2.1) und anschließend die Durchführung (Abschnitt 5.1.2.2).

Tabelle 5.1.: Anbieterauswahl Vollzugsdefizitanalyse

	Gesamt	Nachrichtenportale	Shops	Auktionsplattformen	E-Mail	Messenger	soziale Netzwerkseiten
Anzahl untersuchter Anbieter	100	21	48	8	6	10	7
Anzahl Registrierungen	89	21	38	8	6	9	7

5.1.2.1. Aufbau

In diesem Abschnitt geben wir Informationen zu der Anbieterauswahl. Außerdem beschreiben wir eine virtuelle Identität, die wir bei der Studie einsetzen.

Anbieterauswahl Für unsere Studie haben wir 100 Diensteanbieter im Internet untersucht. Anhang C nennt die Unternehmen namentlich. Um eine Vielfalt interessanter Anbieter miteinzubeziehen, haben wir uns an Anbieterbewertungen zum Marktanteil, Anzahl Klicks [Arb08, EF08] und Rankings nach Kundenzielgruppe [Med06a, Med06b] orientiert. Die Menge der Anbieter umfasst populäre Nachrichtenportale, Shops, Auktionsplattformen, E-Mailanbieter, Betreiber von Systemen zum Austausch von Nachrichten (Instant Messenger) und soziale Netzwerkseiten. Die genaue Verteilung ist in Tabelle 5.1 angegeben.

Drei Kriterien standen bei der Auswahl der Anbieter im Vordergrund:

Einfluss Die Anzahl der Nutzer und Interaktionen der Nutzer mit einem Anbieter definieren den Einfluss eines Anbieters. Unsere Auswahl umfasst Anbieter mit einem großen Marktanteil. Außerdem berücksichtigen wir unterschiedliche Gruppen von Anbietern, jeweils interessant für Personen unterschiedlichen Alters und aus unterschiedlichen sozialen Gruppen.

Relevanz Wir haben die Anbieter aus verschiedenen Kategorien, namentlich Nachrichtenportale, Online Shops, Auktionsplattformen, E-Maildiensten, Messenger-Diensten und sozialen Netzwerkseiten ausgewählt. Diese Kategorien sind relevant für die meisten Internetanwender unserer Gesellschaft. Die Anzahl der Anbieter pro Kategorie ist unterschiedlich, da nicht für jede Kategorie gleich viele relevante Anbieter existieren. Zum Beispiel ist die Anzahl nennenswerter Suchmaschinen deutlich geringer als die Anzahl Shops.

Vergleichbarkeit Die Datenschutzpraktiken der Anbieter sind nur dann wirklich vergleichbar, wenn wir sie auch nach den gleichen Gesetzen bewerten. Wir haben uns für Anbieter entschieden, die nach dem deutschen Recht zu bewerten sind, im Speziellen nach dem Telemediengesetz (TMG) und dem Bundesdatenschutzgesetz (BDSG). Dies schließt jedoch keine international tätigen Unternehmen aus.

Virtuelle Identität Wissen die Anbieter, dass sie Teil einer Studie sind, ist davon auszugehen, dass sie sich anderes verhalten als beim alltäglichen Umgang mit dem Thema Datenschutz. Aus diesem Grund haben wir eine künstliche Identität geschaffen. Diese hat einen Namen, Geburtsdatum, eine reale Adresse, Briefkasten, Telefonnummer, Faxnummer, Mobiltelefon, E-Mail und ein Pseudonym. Wir haben die virtuelle Identität für die Interaktion mit dem Anbieter verwendet.

5.1.2.2. Durchführung

In diesem Abschnitt beschreiben wir die Durchführung der Studie. Diese umfasst die folgenden vier wesentlichen Schritte:

1. Externe Analyse In einem ersten Schritt haben wir die Webseite des Anbieters analysiert. Dies beinhaltet insbesondere die Auffindbarkeit der Datenschutzerklärung und das Prüfen auf die Erfüllung der Informationspflichten.

2. Registrierung Im zweiten Schritt haben wir uns mit unserer künstlichen Identität bei den Anbietern angemeldet und den Registrierungsprozess analysiert. Von besonderem Interesse dabei sind die Daten, die zur Diensterbringung erhoben werden und die in vielen Fällen abzugebende Einwilligung in besondere Datenschutzpraktiken. Die Einwilligung ist immer dann erforderlich, wenn die Praktik über das hinausgeht, was der Gesetzgeber vorsieht (und kein anderes Gesetz die Praktik legitimiert). Bei 11 der 100 Anbieter war dieser Schritt nicht möglich beziehungsweise erforderlich, da der Anbieter zum Beispiel keine Registrierung für den Kauf eines Produktes verlangt hat.

3. Auskunftsersuchen An die 89 Anbieter, bei denen wir uns registrieren konnten, haben wir im Namen unserer künstlichen Identität ein Auskunftsersuchen gestellt. Wir haben (1) nach gespeicherten personenbezogenen Informationen gefragt und (2), ob der Anbieter Informationen weitergegeben hat und wenn ja, an wen. Das Auskunftsersuchen hat wie folgt ausgesehen:

> Gemäß des deutschen Bundesdatenschutzgesetzes und Telemediengesetzes fordere ich Sie auf, mir bis spätestens Tag.Monat Auskunft über folgende Punkte zu geben:
>
> 1. Informationen über alle (personenbezogenen) Daten, die Sie über mich beziehungsweise von mir, gespeichert haben. Bitte geben

Sie die konkreten (i) Attribute und (ii) Attributwerte, sowie (iii) deren Verwendungszweck an.

2. Informationen darüber, (i) welche der Daten Sie an Dritte weitergegeben haben, (ii) den Name und Kontakt jeden Empfängers, sowie (iii) den jeweiligen Verwendungszweck.

Bitte schicken Sie mir die Daten an ...

4. Löschung der Daten Im letzten Schritt unserer Studie haben wir im Namen unserer künstlichen Identität an alle Anbieter eine E-Mail ähnlich obigen Beispiels geschickt. Darin haben wir die sie aufgefordert, die personenbezogenen Daten zu löschen.

5.1.3. Evaluation

Im Folgenden beschreiben wir die Ergebnisse unserer Studie. Diese beziehen sich auf Verstöße in der Datenschutzerklärung 5.1.3.1, bei der Einwilligungserklärung 5.1.3.2, beim Auskunftsersuchen 5.1.3.3 und beim Recht auf die Löschung der personenbezogenen Daten 5.1.3.4.

5.1.3.1. Informationspflichten

Wir haben die Datenschutzerklärungen von 100 Anbietern untersucht. Unser besonderer Fokus hat dabei auf (1) der Abrufbarkeit der Datenschutzerklärung, (2) Informationspflichten bezüglich der Datenerhebung, Verarbeitung und Nutzung, (3) der Datenweitergabe und (4) der automatisierten Datenverarbeitung gelegen. Zu jedem Evaluierungskriterium geben wir den relevanten Paragraphen und sinngemäß dessen Inhalt wieder.

Jederzeitige Abrufbarkeit §13 Abs. 1 Satz 3 TMG: *Ein Kunde muss in der Lage sein, die Datenschutzerklärung einfach und jederzeit abrufen zu können. Der Anbieter sollte zum Beispiel auf jeder Seite einen Link vorsehen, der auf die aktuelle Datenschutzerklärung verweist.*

Tabelle 5.2 zeigt, dass 90 Anbieter Datenschutzerklärungen besitzen, die über farblich hervorgehobene Links von jeder Seite aus einfach zu erreichen sind. Die Datenschutzerklärung ist dabei entweder Teil der Allgemeinen Geschäftsbedingungen (AGBs) oder eine eigenständige Seite. Bei 9 Anbietern muss der Nutzer einer Reihe von Links folgen, bis er zu der Datenschutzerklärung gelangt. In diesem Fall hängt es vom konkreten Aufbau der Webseite ab, ob dieses Verhalten vor Gericht als zulässig oder unzulässig befunden wird. Ein Anbieter hat keine für uns auffindbare Datenschutzerklärung bereitgestellt, verstößt also klar gegen das Gesetz. Für uns interessant, wenngleich auch

Tabelle 5.2.: Abrufbarkeit der Datenschutzerklärung

	Gesamt	Nachrichtenportale	Shops	Auktionsplattformen	E-Mail	Messenger	soziale Netzwerkseiten
Im direkten Zugriff	90	16	45	8	6	8	7
Auffindbar	9	5	2	0	0	2	0
Nicht auffindbar	1	0	1	0	0	0	0
Verweist auf alte Gesetze	10	2	6	0	0	2	0

nicht zwingend ein Verstoß: 10 Anbieter beziehen sich noch auf Gesetze, die Anfang 2007 abgelöst wurden, insbesondere das Teledienstedatenschutzgesetz.

Informierung über Erhebung, Verarbeitung und Nutzung §13 Abs. 1 Satz 1 TMG: *Ein Anbieter muss informieren über (i) die Art der Daten, das heißt, welche Daten (Attribute) erhoben werden, (ii) den Umfang der Daten, insbesondere die Speicherzeit, und (iii) den Zweck der Datenerhebung, Verarbeitung und Nutzung. Ist der Zweck offensichtlich, kann der Anbieter auf dessen Angabe verzichten.*

Art der Daten Die 99 Anbieter, die eine Datenschutzerklärung haben, weisen die bei der Dienstnutzung erhobenen Daten unterschiedlich aus (erster Teil Tabelle 5.3). 47 Anbieter geben explizit an, welche einzelnen Attribute sie speichern. 21 geben grobe, jedoch intuitive Kategorien von Daten an, zum Beispiel 'Lieferadresse'. 31 Anbieter weisen nur unspezifische Formulierungen aus wie 'Daten zur Auftragsabwicklung'. 6 Anbieter machen keine Angaben zu den erhobenen Daten. Während Verstöße bei den unspezifischen Formulierungen wieder vom Einzelfall abhängen, verstoßen zumindest die 6 ohne Angaben klar gegen das Gesetz.

Umfang (Speicherzeit) 68 Anbieter (zweiter Teil Tabelle 5.3) weisen die Speicherzeit aus. Beispielsweise geben sie an, dass Daten gespeichert werden 'bis der Zugang vom Nutzer gelöscht wird'. Einige sperren die Daten nur, da sie aus anderen rechtlichen Verpflichtungen zu deren Aufbewahrung angehalten sind. 31 Anbieter machen keine Angaben zur Speicherzeit, was im Widerspruch zu den gesetzlichen Forderungen steht.

Tabelle 5.3.: Informierung über Datenerhebung und Nutzung

	Gesamt	Nachrichtenportale	Shops	Auktionsplattformen	E-Mail	Messenger	soziale Netzwerkseiten
Art							
Detaillierte Angaben	47	12	18	2	4	5	6
Grobe Kategorien	21	5	10	2	1	2	1
Unspezifische Angaben	31	4	19	4	1	3	0
Speicherzeit							
Daten für eine bestimmte Zeit gespeichert	68	15	26	8	5	8	6
Keine Angabe	31	6	21	0	1	2	1
Zweck der Datenerhebung							
Detaillierte Angaben	78	10	42	7	5	7	7
Unspezifische Angaben	21	11	5	1	1	3	0

Zweck der Datenerhebung 78 Anbieter (dritter Teil Tabelle 5.3) geben den Zweck der Datenerhebung explizit an. 21 Anbieter verwenden unspezifische Formulierungen, wie etwa 'zur Diensterbringung'. Der Gesetzgeber lässt dies jedoch nur dann zu, wenn der Zweck offensichtlich ist. Wir haben jeden Anbieter diesbezüglich analysiert. Bei nur 6 der 21 Anbieter (alles Web-Shops) ist der Zweck offensichtlich. Ist der Versandt eines Produktes erforderlich, ist die Erhebung der Lieferanschrift offensichtlich. Alle anderen 15 Dienste bieten Zusatzfunktionalität, wie zum Beispiel Informationsportale oder E-Maildienste an, oder integrieren Dienste von Drittanbietern. In diesen 15 Fällen haben wir den Zweck als nicht offensichtlich bewertet, das heißt, die Anbieter verstoßen gegen das Gesetz.

Informierung über die Datenweitergabe §13 Abs. 1 Satz 1 TMG: *Jeder Anbieter muss über eine Weitergabe der personenbezogenen Daten unterrichten. Die Datenschutzerklärung muss deutlich machen, welche Daten, an wen, zu welchem Zweck weitergegeben werden.*

64 Anbieter geben in ihrer Datenschutzerklärung an, personenbezogene Daten weiterzugeben (Tabelle 5.4). 23 Anbieter tun dies nach eigener Angabe zur Vertragserfüllung. In vielen Fällen, zum Beispiel innerhalb der EU, ist dies zulässig. 27 Anbieter

Tabelle 5.4.: Informierung über Datenweitergabe

	Gesamt	Nachrichtenportale	Shops	Auktionsplattformen	E-Mail	Messenger	soziale Netzwerkseiten
Anbieter leitet Daten weiter	64	13	30	5	6	7	3
Zweck							
Zur Vertragserfüllung	23	2	18	2	0	1	0
Unspezifischer Grund	27	9	5	0	6	5	2
Empfänger							
Logistikpartner und Kreditkarteninstitute	26	0	24	1	0	1	0
Verbundene Unternehmen	27	7	6	2	5	6	1
...Empfängerunternehmen genannt	7	5	0	2	0	0	0
Unspezifische Angaben zum Partner	22	4	9	0	3	4	2
...Partner werden genannt	1	0	0	0	0	0	1
Empfängerländer							
Nicht EU-Ausland	*12*	0	4	1	3	3	1
...Empfängerländer genannt	*8*	0	2	1	3	1	1

machen jedoch nur unspezifische Angaben, wie ‘um die Dienstqualität zu verbessern’. 26 Anbieter geben an, Daten an Logistikpartner und Kreditkarteninstitute weiterzuleiten. 27 Anbieter leiten personenbezogene Daten an verbundene Unternehmen weiter, davon geben jedoch nur 7 Anbieter diese verbundenen Unternehmen explizit an. Wenn der Empfänger eindeutig genannt wird, kann der Nutzer in vielen Fällen daraus schließen, zu welchem Zweck die Daten weitergegeben werden. Auf der anderen Seite leiten 22 Anbieter Daten an nur vage definierte ‘befreundete Unternehmen’ oder ‘Geschäftspartner’ weiter. Von diesen Anbietern weist auch nur einer die Empfänger explizit aus.

Als letzten Punkt hat unsere Auswertung ergeben, dass 12 Anbieter angeben, die Daten an Unternehmen außerhalb der EU weiterzugeben. In nur 8 dieser Fälle werden die Zielländer genannt, das heißt, in nur 8 Fällen kann ein Nutzer erkennen, welches Datenschutzniveau in diesen Ländern zu erwarten ist. Wenn unspezifische Formulierungen eine Person davon abhalten, seine Rechte, wie das Recht auf Löschung seiner Daten, wahrnehmen zu können, verstoßen die Unternehmen gegen geltendes Datenschutzrecht. Entsprechend kann ein Nutzer, so der Empfänger der Daten gar nicht bekannt ist oder dieser in einem unbekannten Land außerhalb der EU ansässig ist, seine personenbezogenen Daten nicht nachverfolgen und folglich auch niemals löschen – es liegt also ein Verstoß vor.

Informierung über automatisierte Datenverarbeitung §13 Abs. 1 Satz 2 TMG: *Jeder Anbieter muss über den Einsatz automatisierter Verfahren zur Datenverarbeitung unterrichten. Dies gilt insbesondere dann, wenn solche Verfahren die Identifikation einer Person ermöglichen oder unterstützen. Die Informationspflicht umfasst (i) die Art der Daten, (ii) den Umfang (die Speicherzeit) und (iii) den Zweck der Datenerhebung, Verarbeitung und Nutzung.*

Die wohl bekannteste Technologie, die eine automatisierte Datenerhebung und Verarbeitung erlaubt, sind Cookies. Cookies sind kleine Dateien, die auf dem Rechner des Nutzers gespeichert werden und deren Inhalt beim Aufruf einer Seite vom Nutzer an den Anbieter übertragen wird. Der Inhalt der Cookies kann den Zustand eines Einkaufkorbes beschreiben, ein Pseudonym sein etc. Meist beinhaltet die Datei zumindest eine eineindeutige Nummer, die es dem Anbieter erlaubt nachzuvollziehen, wann der Nutzer welche Seiten seines Internetauftritts besucht hat. Es gibt zwei Arten von Cookies: Session Cookies werden automatisch beim Schließen des Browsers gelöscht. Persistente Cookies verbleiben hingegen auf dem Rechner und erlauben oft für lange Zeit das Erstellen von Nutzerprofilen. Ein Extrembeispiel für eine Speicherzeit sind sicherlich 99 Jahre bei der Frankfurter Allgemeinen Zeitung. Da Cookies beim Nutzer gespeichert werden, können wir die Speicherzeit von extern erkennen.

Wir haben untersucht, ob und wie die Anbieter Cookies einsetzen und ob sie die Nutzer über deren Einsatz korrekt informieren.

Während 96 der 100 Anbieter Cookies einsetzen (Tabelle 5.5), so weisen aber nur 72

Tabelle 5.5.: Informierung über automatisierten Datenverarbeitung

	Gesamt	Nachrichtenportale	Shops	Auktionsplattformen	E-Mail	Messenger	soziale Netzwerkseiten
Anbieter mit Cookie-Einsatz	96	21	46	7	6	9	7
Informierung über die automatisierte Datenverarbeitung							
Einsatz von Cookies genannt	72	16	35	2	6	7	6
Zweck der Cookies genannt	65	14	33	1	5	6	6
Speicherzeit der Cookies genannt	28	5	18	0	1	2	2
Rechtliche Aspekte							
Cookie-Einsatz ohne Informierung	24	5	11	5	0	2	1
Keine Informierung über die Speicherzeit	41	11	14	2	5	5	4
Falsche Informierung über die Speicherzeit	9	3	3	0	1	1	1

dieser Anbieter auf deren Einsatz in der Datenschutzerklärung hin. Außerdem nennen nur 65 Anbieter den Zweck des Cookie-Einsatzes und auch nur 28 machen Angaben zur Speicherzeit (zweiter Teil der Tabelle 5.5). Trotz einer Grauzone bei der Bewertung des Cookie-Einsatzes (deshalb addieren sich nicht alle Zahlen zu 100%) können wir sagen, dass 24 Anbieter überhaupt keine Angaben zu Cookies machen. 31 verstoßen gegen das Gesetz, indem sie den Zweck nicht nennen und 41 Anbieter aufgrund fehlender Angaben zu der Speicherzeit der Cookies. Davon geben sogar 9 Anbieter eine falsche Speicherzeit an (dritter Teil der Tabelle 5.5). Nur 19 Anbieter setzen Cookies gesetzeskonform ein.

5.1.3.2. Einwilligung

§12 Abs. 1, §13 Abs. 2 TMG: *Die Erhebung, Verarbeitung und Nutzung personenbezogener Daten ist nur dann gestattet, wenn sie vom Gesetzgeber legitimiert ist, wenn der Zweck der Daten offensichtlich ist, wie bei der Versandtadresse, oder wenn der Nutzer eingewilligt hat. Im Falle einer Einwilligung, muss der Nutzer über sein Recht, die Einwilligung jederzeit widerrufen zu können, informiert werden.*

Gemäß unserer Bewertung müssen 72 Anbieter den Nutzer nach einer Einwilligung fragen (Tabelle 5.6). 12 dieser 71 Anbieter fordern überhaupt keine Einwilligung (Einw.), verstoßen also gegen das Gesetz. 47 erwarten von dem Nutzer, dass dieser

Tabelle 5.6.: Informierung über Einwilligung und Widerruf

	Gesamt	Nachrichtenportale	Shops	Auktionsplattformen	E-Mail	Messenger	soziale Netzwerkseiten
Einw. erforderlich	72	16	27	8	5	9	7
. . . jedoch nicht eingefordert	12	0	10	1	0	1	0
Einw. in die DSE / AGBs gefordert	47	13	8	6	5	8	7
Hinweis auf Recht zum Widerruf	54	16	16	5	4	6	7
Einwilligung in personalisierte Nutzerprofile							
Einw. für Nutzerprofile erforderlich	27	5	6	2	3	6	5
Einw. in DSE oder AGBs gefordert	23	5	2	2	3	6	5
Keine Einw.	4	0	4	0	0	0	0
Einwilligung zur Datenerhebung							
Einw. für Datenerhebung erforderlich	26	2	9	6	1	1	7
Einw. in DSE oder AGBs gefordert	18	2	3	5	1	0	7
Keine Einw.	8	0	6	1	0	1	0

in die vollständige Datenschutzerklärung (DSE) einwilligt, auch wenn viele der dort beschriebenen Praktiken gar keine Einwilligung erfordern. Es hängt vom individuellen Anbieter ab, ob dieses Verhalten zulässig ist oder nicht. Eine Einwilligung ist zum Beispiel dann ungültig, wenn sie in einer riesigen, kaum überschaubaren Erklärung verborgen ist. Eine farbige Hervorhebung der Einwilligungserklärung in zum Beispiel den AGBs ist hingegen zulässig.

Der zweite und dritte Teil von Tabelle 5.6 führt unser Ergebnis im Detail aus. 27 Anbieter erstellen personalisierte Nutzerprofile, vier von ihnen fordern jedoch keine Einwilligung. 23 Anbieter fordern die Einwilligung in einem umfangreichen, schwer erfassbaren Dokument. Das ist je nach Einzelfall ein Verstoß. Ähnlich verhält es sich bei der Erhebung von Daten, die für die Diensterbringung nicht erforderlich sind. 26 Anbieter erheben mehr Daten als erforderlich, 18 von ihnen fordern eine Einwilligung in einem umfangreichen Dokument, 8 fragen gar nicht erst nach einer Einwilligung – ein weiterer Verstoß.

Eine gültige Einwilligung erfordert, dass der Nutzer jederzeit prüfen kann, in was er eingewilligt hat. Insbesondere, wenn der Anbieter seine Datenschutzerklärung geändert hat, kann es sein, dass der Nutzer nur in eine Vorversion der Erklärung eingewilligt hat.

Nur wenige Anbieter stellen wie Amazon frühere Versionen der Datenschutz- beziehungsweise der Einwilligungserklärung bereit. Bei keinem Anbieter haben wir erkennen können, in welche Version wir tatsächlich eingewilligt haben.

5.1.3.3. Auskunftsersuchen

§13 Abs. 7 TMG, §34 BDSG: *Jeder Nutzer kann bei einem Dienstanbieter anfragen, welche personenbezogenen Daten über ihn gespeichert sind. Der Anbieter muss diese Anfrage beantworten und alle gespeicherten Daten auflisten sowie mitteilen, an wen er die Daten weitergegeben hat.*

Tabelle 5.7.: Qualität der Auskunftsersuchen

	Gesamt	Nachrichtenportale	Shops	Auktionsplattformen	E-Mail	Messenger	soziale Netzwerkseiten
Anzahl Antworten	56	14	27	4	2	4	5
Mittlere Antwortzeit	2,1	1,4	2,7	2,5	0,5	3,3	2
Keine Informationen, aber Hinweis auf die Datenschutzerklärung	8	4	1	1	1	1	0
Keine Informationen, aber Hinweis auf die Profilseite	7	3	2	0	1	0	1
Unbrauchbare Antwort	6	2	3	0	0	0	1
Gibt an, keine Daten weitergeleitet zu haben	25	6	11	3	1	0	4
Macht falsche Informationen über die Datenweitergabe	2	2	0	0	0	0	0
Keine Angaben zur Datenweitergabe	16	5	8	0	1	1	1

Wir haben unsere künstliche Identität benutzt, um bei 87 Anbietern anzufragen, (i) welche Daten sie über die künstliche Identität gespeichert haben und (ii) welche Daten sie an andere Unternehmen weitergegeben haben. Wie Tabelle 5.7 zeigt, haben wir 56 Antworten erhalten. 31 Anbieter haben unser Auskunftsersuchen ignoriert, ein klarer Verstoß. Die Anbieter haben im Mittel innerhalb von zwei Tagen geantwortet. Einige Anbieter haben nicht mit den gespeicherten Daten geantwortet, sondern nur angegeben, dass sie die Daten wie in der Datenschutzerklärung genannt, speichern. Unserer Bewer-

Tabelle 5.8.: Qualität der Löschersuchen

	Gesamt	Nachrichtenportale	Shops	Auktionsplattformen	E-Mail	Messenger	soziale Netzwerkseiten
Anzahl Antworten	59	14	31	4	3	2	5
Mittlere Antwortzeit	1.2	1.0	2.0	1.5	1	1	0.4
Sofortige Löschung	35	5	17	4	2	3	4
Selbst Löschung	5	1	2	1	0	0	1
Löschung abgewiesen	2	1	1	0	0	0	0
Nicht registriert	5	1	4	0	0	0	0

tung nach ist das nicht ausreichend, da hier die konkreten Attributwerte und nicht die Attribute erforderlich sind. Ansonsten kann ein Nutzer zum Beispiel nicht sein Recht auf die Korrektur falscher Daten durchsetzen. Insgesamt haben 8 Anbieter die Daten nicht konkret genannt. 7 Anbieter haben uns auf die Profilseite ihres jeweiligen Dienstes verwiesen. Unter der Annahme, dass sie wirklich nur die Daten der Registrierung gespeichert haben, ist dieses Verhalten korrekt. 6 Anbieter haben mit Standardphrasen geantwortet, die nichts mit unserem Auskunftsersuchen zu tun haben, also ebenfalls gegen das Gesetz verstoßen. 25 Anbieter haben angegeben, keine Daten weitergeleitet zu haben. In 2 Fällen kam die Antwort auf unser Auskunftsersuchen von einem anderen Unternehmen als dem angefragten. Somit können wir davon ausgehen, dass der Zugriff auf unsere Daten auch von dem anderen Unternehmen aus möglich war.

5.1.3.4. Löschersuchen

§35 Abs. 2 BDSG, §12 Abs. 1, Abs. 2, §14 Abs. 1 und §15 TMG: *Ein Anbieter muss auf Forderung des Nutzers alle personenbezogenen Daten löschen. Ausnahmen beinhalten rechtliche Pflichten, die Daten weiter zu speichern oder Abrechnungszwecke. In diesen Fällen sind die Daten zu sperren.*

Wir haben im Namen unserer künstlichen Identität eine Forderung auf Löschung unserer personenbezogenen Daten an alle 87 Anbieter geschickt, bei denen wir registriert gewesen sind. Wie aus Tabelle 5.8 ersichtlich, haben 59 Anbieter auf unsere Forderung reagiert, während 28 Anbieter die Löschaufforderung ignoriert haben und somit einen Verstoß begehen. 35 der 59 Anbieter haben die personenbezogenen Daten direkt oder

nach einer kurzen Bitte um Bestätigung gelöscht sowie die Löschung bestätigt. 5 Anbieter haben angegeben, nicht mehr als die Daten zu speichern, die man im Profil selbst entfernen kann. 2 Anbieter haben eine Löschung aus technischen Gründen zurückgewiesen, ein klarer Verstoß. 5 Anbieter haben behauptet, wir seien nicht registriert oder die Daten könnten nicht gefunden werden. Wir haben auch nach unserem Löschersuchen die Zugänge überprüft und haben uns nach wie vor anmelden können. Insgesamt haben 35 Anbieter beim Löschersuchen gegen geltendes Recht verstoßen.

5.1.4. Diskussion und Zusammenfassung

Obige Studie ist eine interdisziplinäre Arbeit, bei der wir die Datenschutzpraktik von 100 Diensteanbietern im Internet untersucht haben. Alle Anbieter sind nach dem Telemediengesetz und dem Bundesdatenschutzgesetz zu bewerten und sind relevant für eine große Gruppe von Internetnutzern. Aufgrund der Harmonisierung des Datenschutzrechts durch die EU gehen wir davon aus, dass die Ergebnisse auch stellvertretend sind für andere EU-Mitgliedsstaaten.

Unsere Ergebnisse zeigen, dass das in Abschnitt 4.3 identifizierte mangelnde Vertrauen von Nutzern in Diensteanbieter berechtigt ist. Die große Mehrheit von Anbietern verwendet in ihren Datenschutzerklärungen unspezifische Formulierungen, aus denen der Nutzer nicht ableiten kann, an wen seine Daten fließen. Die Anbieter informieren die Nutzer nicht oder falsch und reagieren nicht auf die Auskunfts- und Löschersuchen – den aus Nutzersicht wohl wichtigsten Mechanismen im Datenschutz. Wir leiten außerdem aus den zumeist unstrukturierten Antworten der Datenschutzbeauftragten ab, dass diese keine standardisierten Prozesse für Auskunfts- und Löschersuchen haben. Ebenso haben viele Datenschutzbeauftragte keinen Überblick über die im eigenen Unternehmen eingesetzten Informationssysteme. Obwohl wir registriert sind und uns bei ihrem Dienst anmelden können, finden einige Datenschutzbeauftragte unsere Profile nicht. Erschreckend ist auch der Umgang mit Verfahren zur automatisierten Datenverarbeitung. Wir haben nur die prominenteste und zugleich relativ einfach zu identifizierende Technik der Cookies evaluiert. Deutlich schwieriger zu erkennende Techniken wie Web-Statistikwerkzeuge oder Pixel, das heißt nur ein Pixel große, transparente und damit unsichtbare Bilder, sind deutlich schwieriger zu überprüfen. Es ist davon auszugehen, dass bei diesen Techniken die Anzahl der Datenschutzverstöße sogar noch deutlicher ist.

Bei 100 Anbietern haben wir mehr als 300 Verstöße identifiziert – nur fünf Anbieter verhalten sich korrekt. Damit haben wir das Vollzugsdefizit im Datenschutz für Online-Dienste belegt.

5.2. Kollaborative Identifikation von Datenschutzverstößen

5.2.1. Ein möglicher Ausweg aus dem Vollzugsdefizit

In den vergangenen Jahren hat die Presse nahezu jeden Tag neue Meldungen zu Datenschutzverstößen veröffentlicht. Institutionen wie das 'Identity Theft Resource Center'[2] machen nichts anderes, als diese unzähligen Verstöße – und die daraus resultierenden Konsequenzen für die betroffenen Personen – zu erheben und zu protokollieren. Wie wir im vorangegangenen Kapitel 5.1 gezeigt haben, liegt dies unter anderem in dem Vollzugsdefizit im Datenschutz begründet. Gesetze existieren, es kümmert sich jedoch niemand um deren Durchsetzung. Dabei kann davon ausgegangen werden, dass viele der von uns identifizierten Datenschutzverstöße nicht hätten stattfinden müssen – wenn geltendes Recht und die daraus resultierenden Vorgaben durchgesetzt würden.

5.2.1.1. Ursachen für das Vollzugsdefizit

Wir sehen drei grundlegende Ursachen für das Vollzugsdefizit:

Beschränkte Ressourcen Die Ressourcen der Datenschutzaufsichtsbehörden sind beschränkt, während die Anzahl der Daten erhebenden, verarbeitenden, nutzenden und Daten weiterleitenden Dienste ungebrochen (exponentiell) steigt [NS10]. Die Anzahl der registrierten Domains für die Ballungsräume in Deutschland hat nach Angaben der DENIC[3] vom Jahr 2007 zum Jahr 2008 um 12% zugelegt. Heute[4] sind 13.561.026 de-Domains registriert. Laut einem Interview der TAZ[5] zu unserer Vollzugsdefizitanalyse (Kapitel 5.1) mit dem Landesamt für Datenschutz in Schleswig-Holstein, stehen jedoch nur 6 Angestellte der Datenschutzaufsicht für die Überwachung des Internetauftritts von etwa 100.000 Unternehmen zur Verfügung. Bundesweit sind es sogar nur 2,2 Stellen [Xam09].

Expertenwissen Interessierte Personen brauchen juristisches Expertenwissen, um die Datenschutzgesetze zu verstehen und auf eine konkrete Situation anwenden zu können. So müssen beispielsweise Formulierungen wie 'Angemessenheit der Nutzung' ausgelegt werden, das heißt, es muss entschieden werden, ob die Nutzung der Daten im aktuell betrachteten Kontext rechtens ist oder nicht. Darüber hinaus gibt es in Deutschland mehr als 1.000 Datenschutznormen, die selbst ein Experte nur schwierig alle im Blick behalten kann.

Wandel des Datenschutzbeauftragten Unter den existierenden Datenschutzbeauftragten im öffentlichen wie im privatwirtschaftlichen Bereich bedarf es eines Wandels [agi10]. Die Diversifikation von Unternehmen, die Globalisierung und die

[2] http://www.idtheftcenter.org, Breach Database, April 2010

[3] DENIC eG, zentrale Registrierungsstelle für Domains in Deutschland, Zahlen für 2009 stehen aus.

[4] Stand 26. März 2010

[5] http://www.taz.de/1/politik/schwerpunkt-ueberwachung/artikel/1/unternehmen-ist-datenschutz-egal/, September 2009.

regionale Diversifikation sowie die Stellung eines Datenschutzbeauftragten direkt unter der Geschäftsführung [KB10] sind nur ein paar Gründe, warum sich das Bild von einem Datenschutzbeauftragten und dessen Anforderungsprofil ändert. [agi10] fordert mehr Agilität und Kreativität der Datenschutzbeauftragten der Zukunft – mit detaillierten Kenntnissen in Recht, Technik und unternehmensinternem Gebaren.

5.2.1.2. Das Vollzugsdefizit aus unterschiedlicher Perspektive

Das identifizierte Vollzugsdefizit im Datenschutz betrifft alle am Austausch von Daten beteiligte Parteien.

Unternehmen Während wir in Kapitel 4 die Privatheitsproblematik zwischen den Nutzern untersucht haben, liegt in diesem Kapitel die Verantwortung für einen Datenschutzverstoß klar bei den Unternehmen. Wie vereinzelte Rückläufe, zum Beispiel der VZ-Netzwerke, MyMuesli und Freenet, auf die Vollzugsdefizitanalyse (Kapitel 5.1) gezeigt haben, ist das Thema Datenschutz zu einem Reputationsfaktor geworden. Viele Unternehmen versuchen sich an geltendes Recht zu halten – ohne eine einheitliche Kontrollmöglichkeit zu haben, inwieweit ihnen das gelingt.

Betroffene Person Den betroffenen Personen fehlt in den meisten Fällen das juristische Expertenwissen, um entscheiden zu können, ob ein Unternehmen einen Datenschutzverstoß begeht. Und vermuten Nutzer einen Verstoß bei einem Anbieter, fehlt ihnen (ebenso wie den Unternehmen) ein einheitlicher, intuitiver Bewertungsmaßstab, der es dem Nutzer anhand geltenden Rechts erlaubt, seinen Verdacht zu überprüfen.

Behörden Die Behörden zur Überwachung von Datenschutzpraktiken der Unternehmen sind wie erwähnt personell gar nicht in der Lage, geltendes Recht flächendeckend durchzusetzen. Nach [Xam09] wurden selbst erkannte Verstöße wegen Personalmangels nicht verfolgt. Bei 658 Eingaben (ca. 2 Tage Bearbeitungszeit / Eingabe), 2.000 Anrufen (ca. 10 Minuten / Anruf) und 239 Beratungsanfragen (halber Tag / Anfrage) ist das weiter kein Wunder.

Zertifizierungsinstitutionen Man sollte annehmen, dass Zertifizierungsinstitutionen einen Bewertungsmaßstab zur Zertifizierung von Datenschutzpraktiken von Unternehmen entwickelt haben und diesen auch anwenden. Unsere Beobachtungen, zum Beispiel aus Kapitel 3 und Kapitel 5.1, haben jedoch keinen Unterschied zwischen zertifizierten und nicht zertifizierten Unternehmen erkennen lassen.

Gäbe es für die Unternehmen ein Bewertungsschema zur Selbstüberprüfung der Datenschutzpraktik, könnten die Unternehmen unter Umständen viele Verstöße selbst erkennen. Die betroffenen Personen könnten ohne Datenschutzexpertise Datenschutzverstöße identifizieren. Die Behörden könnten zur Voraussetzung machen, dass vor dem Melden eines Verstoßes überprüft werden muss, ob der Verdacht gemäß des Bewer-

tungsschemas vorliegt. Außerdem könnten sie sich auf solche Verstöße konzentrieren, die viele Nutzer identifiziert haben. Die Zertifizierungsinstitutionen könnten ihre Bewertungskriterien vereinheitlichen und somit untereinander vergleichbar werden.

5.2.1.3. Beiträge und Forschungsfragen

In diesem Kapitel stellen wir *CLEF* (Collaborative Law Enforcement Framework) vor, unseren Ansatz, mit dem wir dem Vollzugsdefizit entgegenwirken [5]. Der Ansatz adressiert interessierte Nutzer, Unternehmen, Datenschutzaufsichtsbehörden und Zertifizierungsinstitutionen. Er erlaubt es Nutzern, die Datenschutzpraktiken von Online-Anbietern wie Web-Shops, Foren, Suchmaschinen etc. auf intuitive Weise zu untersuchen. Zu diesem Zweck haben wir den Verstoßkatalog aus Kapitel 5.1 in eine Taxonomie einfacher, datenschutzrelevanter Fragen transformiert. Außerdem haben wir Muster von Antworten identifiziert, die auf Datenschutzverstöße hinweisen. Personen ohne Expertenwissen können diese Fragen einfach beantworten. Die Taxonomie beschreibt dabei Beziehungen zwischen Fragen. Zwei Fragen stehen zueinander in Bezug, wenn (i) eine Frage mehr Detailkenntnisse erfordert als eine andere und (ii) eine Frage nur dann gefragt wird, wenn eine andere auf eine bestimmte Weise beantwortet worden ist. Letztendlich gleicht *CLEF* zur Identifikation von Datenschutzverstößen die Antwortmuster mit den Antworten der Nutzer ab. Der kollaborative Aspekt unseres Ansatzes erlaubt es Personen, die Fragen zu beantworten, die sie beantworten können, und lässt sie gleichzeitig von den Antworten anderer Nutzer profitieren. Wir berücksichtigen, dass einige Verstöße von komplexen Antwortmustern abhängen und dass manche Verstöße auf mehrere verschiedene Arten identifiziert werden können. Beispielsweise können Nutzer den Registrierungsprozess und die Datenschutzerklärung beobachten oder den Anbieter auf Cookies und Web-Statistikwerkzeuge hin untersuchen.

Beispiel 12: Man stelle sich die rechtlichen Aspekte bei der Datenerhebung, der Datenweitergabe und der Erfordernis einer Einwilligung des Nutzers vor. Fragen bezüglich dieser Aspekte umfassen q_1 ='Fragt das Unternehmen nach personenbezogene Daten?', q_2 ='Gibt das Unternehmen Daten an Unternehmen außerhalb des Europäischen Wirtschaftsraums (EWR) weiter?', q_3 ='Werden die Daten an Argentinien, Guernsey, Isle of Man, Kanada oder die Schweiz weitergeleitet?'[6] und q_4 ='Mussten Sie zu diesem Zweck einwilligen?'. Eine Gruppe interessierter Nutzer beantwortet diese Fragen für mehrere Anbieter. Das Antwortmuster `<ja, ja, nein, nein>` modelliert einen Verstoß gegen das Datenschutzrecht für alle EU-Mitgliedsstaaten. Wenn eine signifikante Anzahl von Antworten einem Muster entspricht, weist *CLEF* einen Verstoß aus.

Beispiel 12 zeigt auch, dass die in Kapitel 2.4 vorgestellten Verfahren nicht ausrei-

[6]Die EU listet Länder mit einem gleichwertigem Datenschutzniveau wie das im EWR. Außerdem lassen wir in diesem Beispiel US Unternehmen außen vor, die sich gemäß dem Safe-Harbor-Agreement verhalten.

chend sind. Elektronisch auswertbaren Datenschutzerklärungen in P3P [Mar02] fehlt es an Ausdrucksmächtigkeit. Verfahren, die auf den Klartextdatenschutzerklärungen mit Hilfe von Sprachverarbeitungstechniken (NLP) aufsetzen, sind bisher nicht erfolgreich [BKK06]. Außerdem sind sie wie P3P auf Inhalte der Datenschutzerklärungen beschränkt. Unser Ansatz betrachtet darüber hinaus den Registrierungsprozess, den Einsatz von Techniken zur automatisierten Datenerhebung, die Unterscheidungen zwischen den AGBs, der Datenschutz- und der Einwilligungserklärung, die Art der Einwilligung und vieles mehr. Generatoren von Datenschutzerklärungen können nur ausdrücken, was der Verantwortliche eines Anbieters dem Generator vorgibt, nicht jedoch, was der Anbieter wirklich macht. Verfahren, die kollaborative Aspekte zur Durchsetzung von Datenschutzpräferenzen nutzen, gibt es [Fel07, FE08, NS09]. Unser Ansatz ist jedoch ein deutlich anderer: Unsere Grundlage ist die Implementierung von rechtlicher Expertise in ein Informationssystem. Dieses Informationssystem erlaubt es Nutzern, Datenschutzverstöße systematisch zu identifizieren. Das heißt, dass das Gesetz entscheidet, ob ein Verstoß vorliegt oder nicht. Das steht im Kontrast zu den subjektiven Vorstellungen von Personen über das, was privat sein sollte. Der Vorteil unseres Ansatzes, neben anderen, ist, dass Nutzer, die einen Verstoß mit *CLEF* korrekt identifiziert haben, auch gleich die rechtliche Grundlage erhalten, nach der sie ihr Recht einfordern können.

Wir weisen darauf hin, dass dieser Ansatz Verstöße identifiziert, jedoch nicht garantiert, dass ein Anbieter konform zu geltendem Recht ist. Jedoch ist unser Ansatz der erste, der vielversprechend ist, dem Vollzugsdefizit entgegenzuwirken und den Umgang mit Datenschutzgesetzen auch Personen ohne Expertenwissen zu ermöglichen. Unser Ansatz kann bei jeder Gesetzesänderung oder jedem Gerichtsurteil leicht angepasst und auf neue Kontexte erweitert werden. Er hat das Potential, eine Vielzahl von Datenschutzverstößen mit der Hilfe interessierter Nutzer zu identifizieren, er erlaubt Unternehmen, interne Prozesse zu überprüfen und ermöglicht den Datenschutzaufsichtsbehörden ein wesentlich effizienteres Arbeiten. Entsprechend hoch bewerten wir auch die gesellschaftliche Bedeutung von *CLEF*.

Nach dem konzeptionellen Entwurf unseres Systems liegt die Implementierung, orientiert an existierenden Web 2.0-Applikationen, mehr oder weniger auf der Hand. Wir konzentrieren uns aus diesem Grund auf die konzeptionelle und weniger die technische Beschreibung.

Beiträge Die Beiträge dieses Kapitels sind wie folgt:

- Wir motivieren und beschreiben *CLEF*, unseren Web 2.0-Ansatz, der es einer Gruppe (Community) von Nutzern erlaubt, Datenschutzverstöße kollaborativ zu identifizieren.
- Wir schlagen eine Methodik vor, mit der Datenschutzexperten Gesetze und Normen auf einfache, intuitive Fragen abbilden können. Exemplarisch leiten wir mit

Hilfe dieser Methodik für ausgewählte Verstöße die Taxonomie her und evaluieren unsere Methodik anhand einer Vorstudie.
- Aufbauend auf den Erkenntnissen aus der Vorstudie erstellen wir eine umfassende Fragentaxonomie für das deutsche Datenschutzrecht für Online-Anbieter. Da die EU die Datenschutzgesetzgebung harmonisiert, ist die Taxonomie mit leichten Anpassungen auch in anderen EU-Ländern einsetzbar.
- Um herauszufinden, ob eine Community von Nutzern ohne Expertenwissen mit *CLEF* Datenschutzverstöße finden kann, führen wir eine umfangreiche Nutzerstudie durch.

Forschungsfragen der Nutzerstudie Die Forschungsfragen, die wir mit der Nutzerstudie beantworten möchten sind:

F1 Ist es möglich, Datenschutzpraktiken von Unternehmen anhand einfacher, intuitiver Fragen zu erheben?
F2 Ist es Nutzern unseres Ansatzes möglich, diese Fragen in vertretbarer Zeit zu beantworten?
F3 Können die Fragen von einer größeren Nutzergruppe so einheitlich beantwortet werden, dass unser PET daraus sinnvoll eine Gemeinschaftsmeinung bilden kann?
F4 Gegeben eine Gemeinschaftsmeinung zu einer Frage, entspricht diese Meinung auch der richtigen Antwort? Anders ausgedrückt, lassen sich mit *CLEF* kollaborativ Datenschutzverstöße korrekt identifizieren?
F5 Wie stark ist das Ergebnis, das heißt die Qualität der erkannten Verstöße, abhängig vom Hintergrundwissen der Nutzer?

Die Studie umfasst 77 Teilnehmer unterschiedlicher sozialer Gruppen. Wir zeigen, dass die Teilnehmer 81% der Datenschutzverstöße finden, die auch die Experten identifiziert haben.

Der weitere Verlauf dieses Kapitels ist wie folgt aufgebaut: In Abschnitt 5.2.2 stellen wir den *CLEF*-Ansatz vor. In Abschnitt 5.2.3 evaluieren wir unseren Ansatz mittels einer Nutzerstudie. Abschnitt 5.2.4 fasst unsere Ergebnisse zusammen.

5.2.2. Der *CLEF* Ansatz und eine Vorstudie

Wir beschreiben zuerst die Anforderungen an *CLEF* (Abschnitt 5.2.2.1) und deren Umsetzung (Abschnitt 5.2.2.2). Anschließend stellen wir die Methodik zur Erstellung einer Taxonomie einfacher Fragen aus Gesetzen und Normen vor (Abschnitt 5.2.2.3). Mittels einer Vorstudie evaluieren wir diese Methodik (Abschnitt 5.2.2.4), das heißt, wir untersuchen an einer Auswahl von Fragen, inwieweit Personen ohne Expertenwissen in der Lage sind, diese zu beantworten.

5.2.2.1. Anforderungen

Unser Ziel ist die Entwicklung eines PETs, mit dessen Hilfe wir dem Vollzugsdefizit entgegenwirken können. Es ist wichtig vorwegzunehmen, dass es an dieser Stelle noch nicht darum geht, ein PET so zu entwickeln, dass es Nutzer wie in Kapitel 4 im Alltag begleitet. Wir kommen diesem Zustand zwar sehr nahe, untersuchen aber das grundlegendere Problem, ob Nutzer mit Hilfe von Software überhaupt in der Lage sind, Datenschutzverstöße anhand einer fundierten rechtlichen Grundlage zu erkennen. Somit sind unsere Anforderungen andere als die in Kapitel 4.3 erarbeiteten.

In diesem Absatz stellen wir die Anforderungen an unser PET vor. Wir konzentrieren uns dabei auf die funktionalen Anforderungen. Die Anforderungen an die Skalierbarkeit, Performanz und Robustheit sind identisch mit den Anforderungen an die meisten existierenden Web 2.0-Anwendungen.

Transparenz (AN1) Nutzer unseres PETs müssen in der Lage sein nachzuvollziehen, welche Praktiken eines Anbieters einen Verstoß darstellen. Unser Ziel ist es, Verstöße zu identifizieren, die auch eine rechtliche Grundlage haben und nicht solche Verstöße, die Nutzer aus rein subjektiver Sicht annehmen. Dazu muss das zu entwerfende PET zu jedem identifizierten Verstoß die rechtlichen Zusammenhänge transparent machen.

Unterstützung unterschiedlicher Nutzergruppen (AN2) Wir müssen berücksichtigen, dass unterschiedliche Nutzergruppen unterschiedliche Anforderungen an das System haben. Beispielsweise haben (offensichtlich) Mitarbeiter eines Unternehmens tieferen Einblick in die Datenschutzpraktiken ihres Unternehmens als deren Kunden. Außerdem haben die einzelnen Nutzergruppen unter Umständen eine unterschiedliche Motivation, Datenschutzverstöße zu identifizieren. Betroffene Personen möchten gegebenenfalls andere Kunden warnen und erkannte Verstöße publik machen. Unternehmen hingegen, die sich selbst auf Verstöße untersuchen, möchten das Ergebnis der Bewertung geheimhalten.

Unabhängiges Datenmodell (AN3) Der rechtliche Rahmen ändert sich kontinuierlich, zum Beispiel durch neue Gesetze oder neue Urteile. Die Modellierung des Rechtsrahmens, die Evaluierung von Unternehmen und die Erkennung von Datenschutzverstößen muss unabhängig voneinander gespeichert werden. Solch eine Trennung erlaubt eine leichte Anpassung des Systems bei rechtlichen Neuerungen.

Stochastische Garantien (AN4) Die unterschiedlichen Interessen der Nutzergruppen (AN2) und der sich ändernde Rechtsrahmen (AN3) erfordern die Auseinandersetzung mit veralteten, fehlerhaften oder unzulänglichen Bewertungen. Wir benötigen somit die Möglichkeit, stochastische Methoden anzuwenden, um zu einem erkannten Verstoß dessen Konfidenz zu ermitteln.

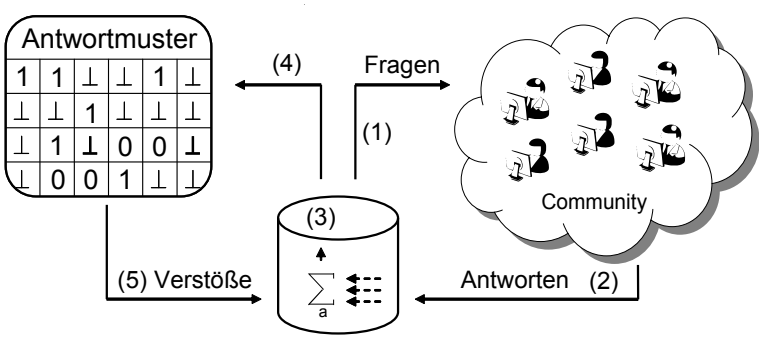


Abbildung 5.1.: Der *CLEF* Ansatz

5.2.2.2. Umsetzung der Anforderungen

In diesem Abschnitt stellen wir unser Systemmodell vor. Zu diesem Zweck führen wir zuerst die Konzepte Fragen, Antworten und Antwortmuster ein. Weitere Informationen zu der Realisierung von *CLEF* befinden sich in Anhang D.1.

Fragen Gesetze beinhalten unspezifische Rechtsbegriffe, Gesetze sind generisch, etc. (siehe Abschnitt 5.2.2.3). Experten transformieren solche Gesetze in konkrete Fragen für spezifische Kontexte, zum Beispiel für Shops im Internet.

Antworten Der Beitrag der Nutzer (-Community) besteht in der Beantwortung dieser Fragen. Die Antworten spiegeln die Erfahrungen und das Wissen der Community über den Anbieter wider.

Antwortmuster Antwortmuster identifizieren anhand bestimmter, durch Experten definierter Kombinationen von Antworten die Datenschutzverstöße.

Abbildung 5.1 zeigt das Prozessmodell von *CLEF*. Experten definieren Fragen, die *CLEF* den Nutzern stellt. Wir sammeln die Antworten (2) und bilden daraus eine Community-Meinung (3). Anschließend prüft *CLEF*, ob die Antworten zu einem der von Experten definierten Antwortmuster passt (4). Wenn ja, so hat unser Ansatz einen Verstoß identifiziert. Verstöße speichern wir in unserer Datenbank (5).

Systemmodell Zur Umsetzung dieses Prozessmodells schlagen wir eine aus drei Komponenten bestehende Architektur vor: Eine Taxonomiekomponente, eine Kollaborationskomponente und eine Komponente zur Erkennung von Verstößen.

Taxonomiekomponente In der Taxonomiekomponente werden die Fragen gespeichert und in einer Taxonomie strukturiert. Unter einer Taxonomie versteht man in der Informationstechnologie eine hierarchische Klassifizierung von Objekten. Die Objekte sind in einer Baumstruktur angeordnet. Eine Beispieltaxonomie ist in Abbildung 5.2 gegeben.

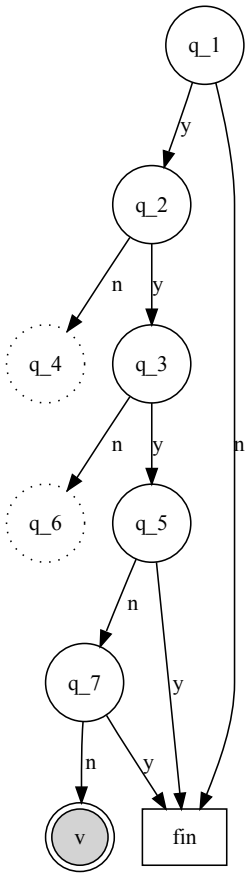


Abbildung 5.2.: Beispiel einer Taxonomie

Die Objekte sind in unserem Fall die Fragen (hier $q_1 - q_7$). Die Annotationen der Kanten in dem Beispiel repräsentieren die Antworten.

Beispiel 13: Antwortet eine Person auf q_1 mit 'nein', so werden keine weiteren Fragen gestellt. Antwortet sie hingegen mit 'ja', so wird ihr Frage q_2 gestellt usw.

Wie an Beispiel 13 ersichtlich, bestehen zwischen den Fragen Beziehungen. Wir unterscheiden zwei Arten von Beziehungen:

Abhängigkeitsbeziehung Abhängigkeiten zwischen Fragen bestehen dann, wenn das Stellen einer Frage nur sinnvoll ist, weil die vorangegangene Frage auf eine bestimmte Art beantwortet wurde.

Abstraktionsbeziehung Eine Abstraktionsbeziehung zwischen Fragen besteht dann, wenn für die Beantwortung zweier Fragen auf unterschiedlichem Abstraktionsniveau unterschiedlich viel Hintergrundwissen erforderlich ist.

Zur weiteren Verdeutlichung das folgende Beispiel:

Beispiel 14: Man stelle sich die Fragen q_1='Setzt der Anbieter Cookies ein?' und q_2='Informiert der Anbieter über die Speicherdauer der Cookies?' vor. Eine Abhängigkeitsbeziehung spezifiziert, dass q_2 nicht gestellt werden muss, wenn q_1='nein' ist. Betrachten wir nun q_3='Nutzt der Anbieter Session-Cookies?'. q_3 ist spezieller als q_1. Eine Abstraktionsbeziehung legt fest, dass q_1 'ja' ist, wenn der Nutzer q_3 mit 'ja' beantwortet.

Wir berücksichtigen in der Taxonomiekomponente (ebenso wie in der später beschriebenen Verstoßkomponente), dass der gleiche Verstoß auf unterschiedliche Art modelliert, aber auch identifiziert werden kann. Varianten der Modellierung können durch eine unterschiedliche Perspektive der Nutzer entstehen, zum Beispiel ob er die Sicht eines Internetnutzers hat oder die eines Mitarbeiters innerhalb eines Unternehmens.

Wir speichern jede Frage als Tupel q:=<QID, Question>Beziehungen zwischen den Fragen speichern wir als r:=<RID, QIDSource, QIDDest, RT.yp>'RID' ist eine eindeutige Bezeichnung für jede Beziehung, 'QIDSource' und 'QIDDest' eine Beziehung und deren Richtung, 'Typ' kann, wie Eingangs dieses Abschnitts beschrieben, die Werte Abstraktionsbeziehung oder Abhängigkeitsbeziehung annehmen.

Kollaborationskomponente Die Kollaborationskomponente verwaltet die Antworten der unterschiedlichen Nutzer. Weiter erlaubt sie unterschiedliche Rollen von Nutzern, nämlich die der Internetnutzer, die der Unternehmen und die der Behörden. Eine besondere, zusätzliche Rolle nehmen die Experten ein, die die Taxonomie erstellen, ändern und erweitern, beziehungsweise die Antwortmuster definieren dürfen.

Nutzer werden gespeichert in der Form u:=<UID, AccountData, Perspect.ive> 'UID' ist die Nutzerkennung, 'AccountData' umfasst Informationen wie den Nutzernamen und optional eine E-Mailadresse, und 'Perspective' ist die Perspektive gemäß der Rolle des Nutzers. Die Kollaborationskomponente stellt die Beziehung zwischen dem Nutzer, der eine Antwort gegeben hat, der Antwort selbst und dem Anbieter, der bewertet wurde, her. Für jede Antwort speichern wir ein Tupel a:=<AID, QID, SID, UID, Answer>Die ID-Felder verweisen jeweils auf die Antwort, die Frage, den Anbieter und den Nutzer. 'Answer' speichert die tatsächliche Antwort.

Aufgrund der unterschiedlichen Perspektiven der Nutzer, eingesetzten Technologien und unterschiedlichen Intentionen der Nutzer, kann es erforderlich sein, dass einzelne Fragen von einer variierenden Anzahl Nutzer beantwortet werden müssen. Die Kollaborationskomponente ermöglicht es, solche Anbieter und Fragen zu identifizieren, die weiteren Nutzern zur Bewertung vorgelegt werden müssen. Außerdem kann sie Nutzer erkennen, die zu Ausreißerantworten neigen (was nicht bedeuten muss, dass eine Ausreißerantwort zwangsläufig falsch ist). Die Kollaborationskomponente erlaubt den Einsatz verschiedener Maße zur Berechnung der Community-Meinung, zum Beispiel eine Gewichtung der Antworten abhängig von der antwortenden Person.

Verstoßkomponente Diese Komponente entscheidet, ob ein Anbieter gemäß der durch die Experten definierten Antwortmuster einen oder mehrere Verstöße begeht. Die Komponente muss dabei vier Aspekte berücksichtigen: *Erstens*, die Qualität der Antworten kann unzureichend sein, um zu einer statistisch signifikanten Aussage zu kommen. *Zweitens* haben die Fragen einen unterschiedlichen Abstraktionsgrad, *drittens* kann der gleiche Verstoß auf mehrere Arten identifiziert werden und *viertens* fordern einige Gesetze, dass k aus n Kriterien erfüllt sind.

Es hätte die Möglichkeit gegeben, komplexe Regelinterpretierer zur Auswertung der Antwortmuster heranzuziehen, mit einem eigenen Datenmodell und einer eigenen Sprache zur Definition der Regeln. Wir haben uns entschieden, einen einfachen Erkennungsmechanismus zu implementieren. Wir modellieren Verstöße als Antwortmuster. Ein Muster ist ein Vektor der Länge '# Fragen' in der Taxonomie. Ein Antwortmuster besteht aus c:=<value, logical operator>-Paaren. Wir speichern ein Muster als pa:=<PAID, QID, value, operator>. Das heißt, jedes Muster hat eine Kennung (PAID) und bezieht sich auf eine oder mehrere Fragen (QID). Der Wert (value) wird über einen Operator (operator) mit den Antworten verglichen. Der Wert in c kann 'ja', 'nein' oder eine Zahl annehmen. Der Operator kann 'nicht relevant' ($\perp$) annehmen, für binäre und kategorische Fragen '=' und für numerische Fragen außerdem $>, <, \leq, \geq$. Der Operator vergleicht den Wert der Antwort mit dem Wert aus dem Muster. Liefert der Operator zu jedem angegebenen Wert eines Musters 'wahr', haben wir einen Verstoß identifiziert.

Beispiel 15: Gegeben die Fragen $q_1 =$'Setzt der Anbieter Cookies ein?' und $q_2 =$ 'Informiert der Anbieter über den Cookie-Einsatz?'. Ein Antwortmuster, das auf einen Verstoß testet, ist $< c_1 := \{wahr, '='\}, c_2 := \{falsch, '='\} >$, wobei alle verbleibenden Werte $\perp$ sind.

Die Muster lassen sich auf einzelne Antworten eines Nutzers anwenden wie auch auf die Antwort, die aus den Bewertungen der Community abgeleitet werden kann.

Alle Elemente eines Vektors sind implizit mit einem logischen UNDverknüpft. Möchte ein Experte mehrere Wege modellieren, über die ein Verstoß erkannt werden soll, so definiert er mehrere Antwortmuster. Diese sind dann automatisch mit einem logischen ODERverknüpft. Einen einmal erkannten Verstoß speichern wir in v:=<VID, UID, PAID, SID, violation>, also mit einer Kennung pro Verstoß (VID), dem Nutzer (UID), der den Verstoß erkannt hat, dem relevanten Antwortmuster (PAID) und dem Anbieter, den der Verstoß betrifft (SID).

Dieser einfache Ansatz hat mehrere Vorteile: Es ist sofort ersichtlich, welche Kombination von Antworten zur Identifikation eine Verstoßes geführt hat (AN1). Die Kollaborationskomponente ermöglicht die Unterscheidung unterschiedlicher Nutzerrollen mit unterschiedlichen Perspektiven (AN2). Die Unterteilung in drei Komponenten (AN3) erlaubt eine einfache Integration neuer Gesetze oder Gerichtsentscheidungen und die relationale Speicherung die Anwendung von etablierten Statistikwerkzeugen (AN4).

5.2.2.3. Methodik zur Erstellung der Taxonomie

Im Folgenden beschreiben wir, wie Experten, zum Beispiel Juristen aus dem Bereich Datenschutz, eine Taxonomie gemäß obiger Beschreibung erstellen können. Die dabei entwickelte Methodik ist das gemeinsame Ergebnis aus der Zusammenarbeit mit dem Lehrstuhl von Prof. Kühling in Regensburg [9].

Die Idee zu der hier entwickelten Methodik ist inspiriert von der allgemeinen Herangehensweise an eine juristische Fragestellung (siehe auch Syllogismus in [Eng05]). Die Ausgangssituation wird dabei durch Fragen wie '*Wer* möchte *was* von *wem* zu *welchem Zweck*?' bestimmt. Zusammen mit Juristen haben wir einen iterativen Prozess entworfen, der es erlaubt, bekannte Sachverhalte, die auf Datenschutzverstöße hinweisen, in Form von Fragen zu modellieren. Insbesondere ermöglicht der Prozess die Formulierung von Fragen, die auch von Nicht-Experten beantwortet werden können.

Der iterative Prozess besteht aus 6 Schritten:

Rechtliche Grundlage Zuerst gilt es zu definieren, welches Gesetz bei dem betrachteten Anbieter Anwendung findet, das heißt, nach welchem Recht der Anbieter zu bewerten ist. In Deutschland ist für die meisten Dienstanbieter im Internet das Telemediengesetz (TMG) beziehungsweise das Bundesdatenschutzgesetz (BDSG) anzuwenden (siehe auch Abschnitt 2.2.2 und Abschnitt 2.2.3).

Beispiel 16: Um zu beantworten, ob das TMG angewendet werden kann, müssen wir von den Nutzern unter anderem wissen q_1 = 'Bietet der Anbieter nur den Internetzugang an?' und q_2 = 'Ist der Dienst ein nicht-moderiertes Web-Forum?'. Wenn q_1 mit 'ja' beantwortet wird, muss das Recht für die Internetzugangsanbieter angewendet werden, nicht das TMG. Ist ein Forum sorgfältig moderiert (q_2), weist das unter Umständen auf einen professionellen journalistischen Hintergrund des Anbieters hin (nach deutschem Recht), das heißt, ebenfalls auf ein anderes Gesetz. Ansonsten kann das TMG angewendet werden.

Perspektive Verstöße können aus unterschiedlicher Perspektive identifiziert werden, zum Beispiel aus dem Unternehmen heraus oder von extern.

Beispiel 17: Man stelle sich q_1 = 'Gibt der Anbieter in der Datenschutzerklärung an, pseudonymisierte Profile zu erstellen?' und q_2 = 'Korreliert der Anbieter die Daten pseudonymisierter Profile mit anderen Datenquellen, so dass eine Identifikation des Betroffenen möglich ist?' vor. Während beide Fragen von einer internen Perspektive aus einem Unternehmen heraus beantwortet werden können, so kann die zweite Frage aus externer Perspektive nicht beantwortet werden, zum Beispiel von einem Internetnutzer.

Konkretisierung Dieser Schritt behandelt unspezifische Formulierungen wie 'zumutbar', 'angemessen', 'sollte' anhand der Gesetzesgrundlage und dem untersuchten Kontext des Anbieters. Zwei Aspekte der Rechtsfolge müssen berücksichtigt werden: (i) Erstens muss untersucht werden, ob gemäß des unspezifischen Rechtsbegriffs ein Datenschutzverstoß vorliegt (Ermessensspielraum), und (ii)

sollte ein Verstoß vorliegen, wie dieser zu ahnden ist (Beurteilungsspielraum).

Beispiel 18: §13 Abs. 6 TMG besagt, dass so lange die anonyme Nutzung eines Dienstes machbar ist, der Anbieter diese Funktion anbieten soll. Im Fall eines Internet-Shops ist es allgemein übliche Praxis, dass ein Kunde den Katalog anonym durchstöbern darf. Das bedeutet, dass solch eine Funktion machbar ist. Anderenfalls begeht der Anbieter unter Umständen einen Verstoß.

Realweltinstanzen Recht ist generisch formuliert. Es verwendet abstrakte Terme, um unterschiedlichste Realweltszenarien, zum Beispiel unterschiedliche Technologien, gleichermaßen zu adressieren. In diesem Schritt wird für den Nicht-Experten die Bedeutung der Gesetzesnorm für unterschiedliche reale Anbieter und Anwendungen in einfache, konkrete Fragen transformiert.

Beispiel 19: Das TMG fordert von einem Anbieter, über den Einsatz von Verfahren zur automatisierten Verarbeitung personenbezogener Daten zu informieren. Es lässt offen, welche Technologien betroffen sind. Die Fragen q_1 = 'Setzt der Anbieter Cookies ein?' und q_2 = 'Verwendet der Anbieter Web-Statistikwerkzeuge?' bilden 'automatisierte Verarbeitung' auf Realweltinstanzen im Internet ab.

Auslegung Auch in diesem Schritt muss mit der Generizität des Rechts umgegangen werden. Bezogen auf die jeweilige Technologie und den Kontext muss die Rechtsfolge herausgearbeitet werden. Dazu gilt es die Wortwahl, die Entstehung und die Absicht (Telos) hinter jeder Norm zu analysieren (vergleiche Abschnitt 2.2.4).

Beispiel 20: Das TMG fordert von einem Anbieter, über die Speicherzeiten von Cookies zu informieren. Es weist jedoch nicht darauf hin, dass die genannte Speicherzeit korrekt sein muss. Trotzdem stellt eine falsche genannte Zeit ein Vergehen da, das modelliert werden muss, da es die Intention des Gesetzgebers war, Transparenz zu schaffen.

Implementierung In diesem Schritt werden die aus den vorangegangenen Schritten entstandenen Fragen zusammengefasst. Die Fragen werden so weit wie möglich vereinfacht, zum Beispiel auf einfache Ja / Nein-Fragen. Es werden Duplikate der Fragen entfernt und die Verstoßmuster definiert.

Dieser Prozess ist iterativ anzuwenden, bis alle Sachverhalte soweit reduziert sind, dass daraus einfache Fragen abgeleitet werden können.

5.2.2.4. Vorstudie

Ziel der Vorstudie ist es zu untersuchen, inwieweit die entworfene Methodik zur Herleitung von intuitiven Fragen, die die Identifikation von Datenschutzverstößen durch Nicht-Experten erlaubt, funktioniert. Weiter ist es unser Ziel, Erfahrungen für die Umsetzung unseres Ansatzes in ein Informationssystem zu erlangen.

Wir werden im Folgenden die Methodik der Vorstudie beschreiben und die Ergebnisse präsentieren. Abschließend diskutieren wir unsere Erkenntnisse, berichten von

Tabelle 5.9.: Beispielfragen aus der Taxonomie der Vorstudie

Fragen ID	Frage
q_1	Setzt der Anbieter Cookies ein?
$q_{1.1}$	Setzt der Anbieter Session-Cookies ein?
$q_{1.2}$	Setzt der Anbieter persistente Cookies ein?
$q_{1.2.1}$	Wie lange sind die persistenten Cookies gültig?
q_7	Weist der Anbieter auf den Einsatz von Cookies hin?
$q_{7.1}$	Informiert der Anbieter über den Zweck des Cookie-Einsatzes?
$q_{7.2}$	... über den Einsatz von Session-Cookies?
$q_{7.3}$	... über den Einsatz von persistenten Cookies?
$q_{7.3.1}$	Welche Speicherzeit für persistente Cookies gibt der Anbieter an?

gewonnenen Erfahrungen im Hinblick auf die Hauptstudie und ziehen ein Fazit (Abschnitt 5.2.2.5).

Methodik In diesem Abschnitt beschreiben wir die Methodik der Vorstudie. Das umfasst Merkmale des Aufbaus der Vorstudie, die gestellten Fragen sowie die untersuchten Verstöße und die Durchführung.

Teilnehmer Unsere Teilnehmergruppe hat aus 27 Studenten bestanden, alle mit einem technischen Studienschwerpunkt. Alle Teilnehmer sind also mit dem Umgang von Online-Diensten vertraut (sie brauchen dieses Wissen zur Anmeldung an Klausuren, dem Kauf von Skripten, Büchern etc.), sie haben jedoch kein juristisches Fachwissen oder Fachwissen im Bereich Datenschutz.

Tabelle 5.10.: Beispielantwortmuster der Vorstudie

Verstoß	q_1	$q_{1.1}$	$q_{1.2}$	$q_{1.2.1}$	q_2	$q_{2.1}$	$q_{2.2}$	$q_{2.3}$	$q_{2.3.1}$
v_1 Keine Information über den Einsatz automatisierter Verfahren.									
	j	$\perp$	$\perp$	$\perp$	j	$\perp$	$\perp$	$\perp$	$\perp$
v_2 Keine Information über den Zweck der eingesetzten automatisierten Verfahren.									
	j	$\perp$	$\perp$	$\perp$	$\perp$	j	$\perp$	$\perp$	$\perp$
v_3 Keine Information über die Cookie-Speicherzeit.									
	$\perp$	j	n	$\perp$	$\perp$	$\perp$	n	$\perp$	`empty`
	$\perp$	$\perp$	j	$\perp$	$\perp$	$\perp$	$\perp$	$\perp$	`empty`
v_4 Falsche Cookie-Speicherzeit.									
	j	$\perp$	$\perp$	$\perp$	$\perp$	$\perp$	$\perp$	$\perp$	$<> q_{1.2.1}$

Fragen Unsere Taxonomie für die Vorstudie umfasst 16 Fragen. Ein Ausschnitt der Fragen ist in Tabelle 5.9 dargestellt. Die Indizes beschreiben Beziehungen zwischen den Fragen. Frage $q_{1.1}$ ist in der Taxonomie unter q_1 angeordnet etc. Wie man sehen kann, handelt es sich sowohl um Abhängigkeitsbeziehungen als auch um Abstraktionsbeziehungen. Die Frage nach der Speicherzeit macht keinen Sinn, so der Anbieter nicht über den Einsatz von Cookies informiert. Die Frage nach persistenten Cookies ist konkreter als nach dem generellen Einsatz von Cookies. Wie an dem Beispiel weiter erkennbar ist, fragen wir sowohl die Informationspflichten ab (q_7 und untergeordnete Fragen), als auch die tatsächliche Praktik des Anbieters, hier bezogen auf den Einsatz von Cookies (q_1 und untergeordnete Fragen).

Verstöße Wir haben für die Vorstudie 6 Verstöße modelliert: (1) Erhebung nicht erforderlicher Attribute ohne Einwilligung, (2) kein Hinweis auf den Einsatz automatisierter Verfahren, oder deren Zweck (3), keine (4) oder fehlerhafte (5) Informationen bezüglich der Speicherzeit erhobener Daten und fehlende Hinweise auf das Recht, seine Einwilligung zu widerrufen (6). Ein Auszug der Verstöße bezüglich der automatisierten Verfahren findet sich in Tabelle 5.10. 'j' bedeutet 'ja', 'n' nein. Ein Verstoß liegt genau dann vor, wenn die Antworten der Nutzer mit dem Antwortmuster (Spalte 2 – 9) identisch sind oder im Antwortmuster $\perp$ für 'nicht relevant' steht. Für v_3 führen zwei unterschiedliche Muster zum gleichen Verstoß, erstens, wenn der Anbieter Session-Cookies einsetzt und zweitens für persistente Cookies. Konkreter bewerten wir es nicht als Verstoß, wenn der Anbieter Session-Cookies und persistente Cookies einsetzt und nur für die persistenten Cookies eine Speicherzeit ausweist.

Anbieter Die Anbieterauswahl für die Vorstudie ist anhand der Ergebnisse aus [6] erfolgt. Orientiert an den Fragen und den modellierten Verstößen haben wir die Anbieter so ausgewählt, dass kein Verstoß bei jedem Anbieter vorkommt, jeder Verstoß jedoch mindestens bei einem Anbieter.

Durchführung Die Durchführung der Vorstudie besteht aus zwei Teilen: *Erstens* einer zehnminütigen Einführung und *zweitens* einer dreißigminütigen Phase zur Bewertung der Anbieter.

Evaluation Die 27 Teilnehmer der Vorstudie haben in der Experimentzeit zwischen 3 und 5 Anbieter bewertet, im Mittel 4,37 Anbieter. In Summe haben wir somit 118 vollständige Bewertungen erhalten mit insgesamt 1816 Antworten. Jeder Anbieter ist von 5 bis 8 Teilnehmern bewertet worden, im Mittel von 5,9 Teilnehmern.

Die Intention der Vorstudie ist die Evaluation unserer Methodik zum Erstellen einfacher, intuitiver Fragen zur Identifikation von Datenschutzverstößen gewesen. Zu diesem Zweck haben wir untersucht, ob (F1) Personen ohne Expertenwissen unsere Fra-

Tabelle 5.11.: Übereinstimmung der Nutzerantworten pro Anbieter (Vorstudie)

κ_{appa}	$\kappa < 0$	$0 \leq \kappa < 0.20$	$0.20 \leq \kappa < 0.40$	
# Anbieter	0	0	2	$\sum = 2$
κ_{appa}	$0.40 \leq \kappa < 0.60$	$0.60 \leq \kappa < 0.80$	$0.80 \leq \kappa < 1.00$	
# Anbieter	4	13	1	$\sum = 18$

gen beantworten können, ob sie dies möglichst einheitlich (F3) tun und ob sich anhand der Antworten Datenschutzverstöße identifizieren lassen (F4). Im Folgenden untersuchen wir zu diesem Zweck zuerst den Grad der Übereinstimmung der Antworten unter den Personen und anschließend, ob die Antworten korrekt sind.

Grad der Übereinstimmung zwischen den Teilnehmern Um den Grad der Übereinstimmung der Antworten zwischen den Teilnehmern zu messen, nutzen wir Fleiss' Kappa [JL.71]. Es erlaubt die Berechnung der Übereinstimmung zwischen mehreren Teilnehmern in Bezug auf mehrere Objekte, hier die Fragen. Nach [LK77] bedeuten Kappa-Werte kleiner 0 keine Übereinstimmung, zwischen 0 und 0,2 eine geringfügige Übereinstimmung, zwischen 0,2 und 0,4 eine mittelmäßige Übereinstimmung, zwischen 0,4 und 0,6 eine angemessene Übereinstimmung, zwischen 0,6 und 0,8 eine beachtliche und zwischen 0,8 und 1 eine perfekte Übereinstimmung. Wir berechnen die Übereinstimmung der Antworten pro untersuchtem Anbieter.

Tabelle 5.11 zeigt unsere Ergebnisse. Für 70% der Anbieter haben wir eine bemerkenswerte bis perfekte Übereinstimmung gemessen. Bei 4 Anbietern hat eine angemessene und bei 2 Anbietern eine mittelmäßige Übereinstimmung vorgelegen.

Richtigkeit der Antworten Die Übereinstimmung der Bewertungen zwischen den Nutzern ist wichtig, um die Intuitivität der Fragen bewerten zu können. Das bedeutet jedoch nicht automatisch, dass die gemeinsame Antwort auch korrekt sein muss. Zu diesem Zweck haben wir zwei Untersuchungen durchgeführt. In einem ersten Schritt haben wir evaluiert, ob die Antworten zufällig zustande gekommen sein könnten. Anschließend haben wir für die modellierten Verstöße gemessen, welche Verstöße die Nutzer richtig beziehungsweise falsch identifiziert haben.

Mit Hilfe eines Binomialtests haben wir für jede Frage geprüft, ob die Null-Hypothese H_0 = 'Übereinstimmung der Antwort ist zufällig' zurückgewiesen werden kann. Die Ergebnisse finden sich in Tabelle 5.12. n gibt die Anzahl der Antworten pro Frage an. Ein Signifikanztest für das Signifikanzniveau von 0,01 (Spalte $sn = 0,01$) hat für alle Fragen ergeben, dass H_0 zurückgewiesen werden kann, die Übereinstimmung also nicht zufällig entstanden ist. Für das Signifikanzniveau von 0,05 haben wir weiter untersucht, was die maximale Wahrscheinlichkeit ist, so dass H_0 gerade noch zurückgewiesen werden kann (Spalte $sn = 0,05$). Wir haben sn von 0,01 auf 0,05 vergrößert,

Tabelle 5.12.: Binomialtest für die Nutzer-Experten-Übereinstimmung (Vorstudie)

Q	n	$sn = 0.01$	$sn = 0.05$	Q	n	$sn = 0.01$	$sn = 0.05$
q_1	118	≈ 0	0.91	$q_{1.1}$	116	≈ 0	0.85
$q_{1.2}$	113	≈ 0	0.85	$q_{1.2.1}$	116	≈ 0	0.88
q_2	115	≈ 0	0.80	q_3	118	≈ 0	0.80
q_4	117	≈ 0	0.90	q_5	117	≈ 0	0.66
q_6	118	≈ 0	0.91	q_7	116	≈ 0	0.91
$q_{7.1}$	115	≈ 0	0.76	$q_{7.2}$	118	≈ 0	0.78
$q_{7.3}$	118	≈ 0	0.78	$q_{7.3.1}$	117	≈ 0	0.89
q_8	116	≈ 0	0.63				

um besser unterscheiden zu können, für welche Fragen H_0 früher und für welche H_0 später zurückgewiesen wird. Diese Aussage lässt fundierte Rückschlüsse zu, welche Fragen schwierig zu beantworten gewesen sind.

Aufsteigend sortiert hat es die größte Unsicherheit bei den Fragen q_5 und q_8 gegeben. Um zu erkennen, warum diese Fragen so unterschiedlich beantwortet worden sind, haben wir uns diese näher angesehen. q_5 fragt, ob die Datenschutzpraktiken in den AGBs erklärt sind. Eigentlich eine einfache Frage. Für die Teilnehmer war sie jedoch nicht konkret genug, da einige der Anbieter Auszüge der eigentlich eigenständigen Datenschutzerklärung nochmals in den AGBs aufführen und umgekehrt in den AGBs Datenschutzpraktiken beschrieben haben, die sie nicht in der Datenschutzerklärung erwähnen. q_8 fragt nach der Informationspflicht des Anbieters, auf das Widerrufsrecht bei einer Einwilligung hinzuweisen. Die Teilnehmer haben hier teilweise Hinweise auf das Löschen eines Accounts als solch einen Hinweis eingestuft.

Zwischenfazit: Zusammenfassend erlauben unsere Ergebnisse zur Grad der Richtigkeit der Antworten und der Übereinstimmung der Antworten die Aussage, dass wir einfache, intuitive Fragen formulieren können, die Datenschutzsachverhalte von Nicht-Experten erkennen lassen (F1). Die gemäß unserer Methodik entwickelten Fragen sind großteils verständlich, und wir können komplizierte, das heißt uneinheitlich beantwortete Fragen, erfolgreich erkennen und verbessern.

Weiter haben wir untersucht, welche Auswirkungen die Antworten der Teilnehmer auf die Identifikation der Verstöße hat (Tabelle 5.13). Das x gibt dabei die korrekte Antwort an und r die Anzahl der Bewertungen je Anbieter s. Die Zahlen in den Spalten der einzelnen Verstöße geben die Anzahl der Teilnehmer an, die den jeweiligen Verstoß identifiziert haben. Ein Verstoß v_i ist von der Mehrheit der Teilnehmer korrekt identifiziert, wenn die Anzahl der Teilnehmer, die den Verstoß erkannt haben, dividiert durch die Gesamtzahl der Teilnehmer, die den Anbieter bewertet haben, größer 50% ist. Bei

Tabelle 5.13.: Identifizierte Datenschutzverstöße (Vorstudie)

S	$\|r\|$	v_1	v_2	v_3	v_4	v_5	v_6
s_1	6		X(6)	X(6)	X(6)		
s_2	6		(2)	(2)	X(6)		
s_3	5		X(5)	X(5)	X(5)		
s_4	6			X(3)	X(6)		
s_5	7				X(6)		
s_6	6				X(6)		(1)
s_7	6	X(2)		X(5)	X(6)		
s_8	7			(2)	X(7)		
s_9	6			(2)	X(6)		(1)
s_{10}	7	X(4)	(1)	X(6)	X(7)		
s_{11}	5			(1)	X(5)		
s_{12}	6				(2)	X(4)	
s_{13}	6				X(5)		
s_{14}	6				X(6)		(1)
s_{15}	8		(1)	(2)	X(7)	(1)	(4)
s_{16}	5		(1)	(1)	X(5)		X(5)
s_{17}	4				X(4)		X
s_{18}	5			(3)	(1)	X(4)	(1)
s_{19}	6				(1)	X(5)	
s_{20}	6			(1)	X(3)	(3)	X(4)

Anbieter s_2 ist beispielsweise Verstoß v_4 zu 100% erkannt worden, gleichzeitig wurden v_2 und v_3 zu $\frac{2}{6}$ fälschlicherweise erkannt. v_4 wurde bei s_{20} nur zu 50% korrekt erkannt – da wir die Mehrheitsmeinung betrachten, bewerten wir diesen Verstoß als nicht erkannt. Im Gegegsatz dazu ist Verstoß v_4 bei s_2 richtig erkannt worden, v_2 und v_3 wurde mehrheitlich nicht falsch ($\neq$ richtig) erkannt.

Tabelle 5.14.: Qualität der Verstoßerkennung (Vorstudie)

Testbedingung	Übereinstimmungen	False Positives	Misses
Einzelne Bewertung	31	22	1
Mehrheitsentscheid	28	4	0

Tabelle 5.14 fasst Tabelle 5.13 zusammen. Sie zeigt, wie viele Bewertungen die Teilnehmer in Übereinstimmung mit der richtigen Antwort abgegeben haben, wie viele nicht (Misses) und wie viele fälschlicherweise erkannt wurden (False Positives). Der Effekt der Community wird durch den Vergleich der Bewertungen einzelner Teilnehmer (erste Zeile der Tabelle) und der Mehrheitsentscheidung deutlich. Wäre jeder der

Bewertungen einzeln einer Datenschutzbehörde zur Nachverfolgung übermittelt worden, so hätte die Behörde die 53 (31 + 22) Bewertungen der 27 Teilnehmer prüfen müssen. Dabei wären zwar 31 Verstöße korrekt erkannt worden, 22 Bewertungen jedoch falsch gewesen, 1 Verstoß wäre nicht erkannt worden. Betrachtet man nur die Verstöße, die jeweils von der Mehrheit der Teilnehmer identifiziert worden sind, so wären 28 Verstöße korrekt erkannt worden. Besonders interessant ist, dass die Anzahl der False Positives von 22 auf 4 reduziert würde.

5.2.2.5. Fazit

Unser wichtigstes Ergebnis und zugleich die Intention dieser Vorstudie zeigt: Es ist möglich, Fragen zum Datenschutz so zu stellen, dass diese auch von Nicht-Experten beantwortet werden können (F1). Insbesondere ist es unterschiedlichen Nutzern möglich, die Fragen homogen zu beantworten (F3). Weiter haben wir zeigen können, dass Personen beim Lesen der Datenschutzerklärungen oder bei der Nutzung eines Dienstes leicht etwas übersehen oder falsch interpretieren, das heißt einzelne Fragen falsch beantworten. Eine Bewertung von 5 bis 7 Personen hat jedoch bereits dazu geführt, dass durch die Kollaboration der Nutzer die große Mehrheit (88%) der tatsächlich vorliegenden Verstöße korrekt erkannt wird (F4). Wichtig dabei ist, auch die Zahl der falsch erkannten Verstöße wird um den Faktor 5,5 kleiner. Aus Sicht einer Behörde, die den Verstößen nachgehen soll, bedeutet das eine immense Zeitersparnis. Die Anzahl der Bewertungen pro Anbieter hat sich als zielführend erwiesen. Wir streben diese auch in der Hauptstudie an.

Die Aufgabe der Hauptstudie wird es sein zu untersuchen, ob die korrekte Identifikation von Datenschutzverstößen auch noch bei einem umfassenden Fragen- sowie Verstoßkatalog möglich ist. Zusätzlich soll sie beantworten, ob dies auch mit unserer Implementierung möglich sowie effizient ist.

5.2.3. *CLEF* Nutzerstudie (Hauptstudie)

Nach der Vorstudie gehen wir davon aus, dass einfache, intuitive Fragen die Identifikation einer Vielzahl von Datenschutzverstößen erlauben.

In diesem Abschnitt beschreiben wir unsere Hauptstudie zur Beantwortung der Forschungsfragen (F1–F5). Die wesentlichen Unterschiede zur Vorstudie sind: (i) Die Taxonomie der Fragen ist deutlich größer und bildet (ii) eine umfassende Menge extern erkennbarer Datenschutzverstöße ab. Eine Übersicht über die verwendete Taxonomie liefert Anhang D.2. Details zu den Fragen und den Verstößen befinden sich in Anhang D.3 respektive Anhang D.4. Weiter (iii) ist auch die Anzahl der Teilnehmer deutlich größer, und die Teilnehmer gehören (iv) Gruppen mit unterschiedlichen rechtlichen wie technischen Vorkenntnissen an. Außerdem (v) haben wir die juristische Expertise, das heißt die Taxonomie und die Antwortmuster, die auf einen Verstoß hinweisen, in ei-

ne Web 2.0-Anwendung implementiert. Der relevante rechtliche Rahmen ist wie in der Vorstudie das Telemediengesetz (TMG) und das Bundesdatenschutzgesetz (BDSG).

Wir stellen zuerst die Methodik unserer Studie vor (Abschnitt 5.2.3.1), gefolgt von unserer Evaluierung (Abschnitt 5.2.3.2) und einer abschließenden Diskussion (Abschnitt 5.2.3.3).

5.2.3.1. Methodik der Hauptstudie

Dieser Abschnitt beschreibt unsere Methodik. Zuerst stellen wir unsere Entwurfsentscheidungen vor, anschließend die Durchführung.

Wesentliche Entwurfsentscheidungen

Teilnehmer Für die Studie haben wir uns für drei Teilnehmergruppen entschieden: (i) Schüler eines Gymnasiums (pup), (ii) Studenten im Hauptstudium technischer Disziplinen (cs) und (iii) Studenten der Rechtswissenschaften (law). Diese haben unterschiedliche Ausbildungen, Erfahrungen mit dem Internet und Interessen. Insgesamt haben 77 Teilnehmer zwischen 13 und 29 Jahren (Mittelwert 22 Jahre) an der Studie teilgenommen. 49 Teilnehmer waren männlich, 28 weiblich.

Perspektive Alle Teilnehmer haben die Rolle eines Internetnutzers eingenommen, der Webseiten von extern betrachtet. Das heißt, die Teilnehmer können Einblick in die Datenschutzerklärung, die AGBs und das Impressum nehmen. Außerdem können sie sich bei dem Anbieter registrieren und verwendete Verfahren zur automatisierten Verarbeitung von Daten, wie zum Beispiel Cookies, Pixel, Web-Statistikwerkzeuge, etc. identifizieren. Sie haben keinen Einblick in Interna des Anbieters.

Anreizmechanismus Üblich bei Nutzerexperimenten ist die Definition eines Anreizes zur Teilnahme. Wichtig für unser Experiment ist, dass dieser Anreiz nicht dazu führt, dass die Teilnehmer (i) möglichst schnell und dabei ungenau Bewertungen von Anbietern vornehmen und (ii), dass keine Motivation entsteht, möglichst viele vermeintliche Verstöße zu erkennen, obwohl diese gegebenenfalls gar nicht vorliegen. Wie haben uns aus diesem Grund für eine pauschale Auszahlung von 20EUR pro Teilnehmer entschieden. Mit einer geplanten Durchführungsdauer von 2 Stunden liegt dieser Betrag in etwa bei dem Gehalt eines Hilfswissenschaftlers.

Anbieter Für die Evaluierung unseres Ansatzes ziehen wir 30 Anbieter aus [6] heran. Die Anbieter sind zumeist große Unternehmen und haben einen großen Marktanteil in

ihrer jeweiligen Domäne. Sie begehen eine Reihe unterschiedlicher Verstöße, die vielfach schwierig zu erkennen sind. Anbieter, die Verstöße wie 'keine Datenschutzerklärung' begehen, die man auch ohne *CLEF* leicht feststellen kann, haben wir aussortiert. Wir haben die Anbieter den Teilnehmern so zugeordnet, dass (i) zwei Teilnehmer immer eine möglichst kleine Menge Anbieter gemeinsam bewerten. Außerdem (ii) haben wir, sobald ein Teilnehmer einen Anbieter fertig bewertet hat, einen neuen Anbieter geladen. Die Teilnehmer haben also über die gesamte Experimentdauer Anbieter bewertet und keinen Vorteil durch besonders schnelles und dadurch gegebenenfalls unpräzises Beantworten der Fragen gehabt.

Verstöße Wir haben 31 Verstöße modelliert, die sich durch die Beantwortung von 43 Fragen aus der Taxonomie ergeben. Es ist uns gelungen, alle Fragen so zu formulieren, dass Nutzer sie mit 'Ja' / 'Nein' beantworten können. Für Fragen, die ein Nutzer nicht beantworten kann, bieten wir darüber hinaus die Antwort 'weiß nicht' an. Zum einfacheren Verständnis haben wir die Verstöße in sechs Kategorien unterteilt: Verstöße (1) in der Datenschutzerklärung, (2) der Datenerhebung und Registrierung, (3) bei dem Einsatz automatisierter Verfahren, (4) bei der Einwilligung, (5) bei pseudonymen Profilen und (6) bei personalisierten Profilen.

Gold-Standard Die Qualität unseres Ansatzes hängt davon ab, ob die Teilnehmer Datenschutzverstöße einheitlich erkennen und ob diese einheitliche Meinung überhaupt korrekt ist. Zu diesem Zweck haben wir vier Experten definieren lassen, welche Verstöße bei den Anbietern vorliegen und diese Bewertung dann als Referenzmaß (Gold-Standard) für die Meinung der Teilnehmer herangezogen.

Der Gold-Standard umfasst für die 30 Anbieter 172 Vorkommen der 31 modellierten Verstöße. Die Anzahl der Verstöße ist relativ groß, jedoch nicht ungewöhnlich (vergleiche Kapitel 5.1). Verstärkt wird der Effekt dadurch, dass wir ähnliche (jedoch ungleiche) Verstöße unabhängig zählen. Beispielsweise sind 'keine Information über den Einsatz von Cookies' und 'keine Information über den Einsatz von Web-Statistikwerkzeuge' Verstöße gegen die Informationspflicht bei automatisierten Verfahren. Um der Anforderung nach Transparenz nachzukommen (AN1), zählen wir diese jedoch einzeln.

Durchführung Die Durchführung unserer Studie hat für jede Nutzergruppe im Labor stattgefunden. Für die Schüler und die Teilnehmer aus den technischen Studiengängen in Karlsruhe, für die Teilnehmer aus dem juristischen Umfeld in Regensburg.

Die Durchführung ist unterteilt in drei Abschnitte:

Einführung (15 Minuten) Zu Beginn haben alle Teilnehmer die gleiche Einführung bekommen. Dies hat eine kurze Beschreibung der Idee des Experiments beinhaltet, den Ablauf der Studie und die Bedienung von *CLEF*.

Anbieterbewertung (140 Minuten) In dieser Phase hat die eigentliche Bewertung der Anbieter mit Hilfe von *CLEF* stattgefunden.

Abschluss (5 Minuten) Zum Abschluss der Studie hat jeder Teilnehmer eine Liste der von ihm identifizierten Verstöße ausgehändigt bekommen. Außerdem haben wir einen Fragebogen ausgegeben, mit dem wir demographische Daten erhoben und Kontrollfragen zum Verständnis des Systems gestellt haben.

5.2.3.2. Evaluation der Hauptstudie

In diesem Abschnitt stellen wir die Ergebnisse vor. Wir orientieren uns in der Darstellung an der Reihenfolge der Forschungsfragen. Das heißt, wir untersuchen zuerst, in welcher Zeit die Bewertungen vorgenommen werden können, wie hoch der Grad der Übereinstimmung der Antworten unter den Teilnehmern ist und zuletzt, inwieweit die Meinung der Teilnehmer mit der korrekten Antwort übereinstimmt.

Zeit für die Bewertung Die 77 Teilnehmer unserer Studie haben in Summe 11.138 Fragen beantwortet, im Mittel also 145 Fragen pro Person. 3,8% der Fragen haben die Teilnehmer nicht beantworten können, anders ausgedrückt haben sie 10.711 Fragen klar mit 'Ja' / 'Nein' beantwortet. Im Mittel hat die Zeit der Studie zur vollständigen Bewertung aller Antworten bei 4,18 Anbietern gereicht.

CLEF stellt manche Fragen abhängig von Antworten auf zuvor gestellte Fragen. Kurz, die Anzahl der zu beantwortenden Fragen kann pro Anbieter variieren. Im Mittel sind für eine vollständige Bewertung 31,8 Fragen notwendig gewesen. Die Teilnehmer haben dazu, ebenfalls im Mittel, 22 Minuten gebraucht. Das stimmt mit Ergebnissen aus anderen Studien [MC08a] überein, bei der Teilnehmer zwischen 18 und 26 Minuten zum Lesen und Verstehen einer Datenschutzerklärung brauchen.

Weiter hat uns interessiert, inwieweit eine Gewöhnung an *CLEF*, das heißt ein Lerneffekt, die gemessene Zeit beeinflusst. Die meisten Teilnehmer haben genau 4 Anbieter bewertet. Für diese Teilnehmergruppe haben wir untersucht, wie sich der zeitliche Bedarf im Mittel vom ersten, zweiten, dritten und vierten Anbieter verändert. Zusammengefasst brauchten nahezu alle Teilnehmer für den ersten Anbieter länger, im Mittel 26 Minuten. Bereits beim dritten und vierten Anbieter ist die Zeit im Mittel konstant bei 21 Minuten geblieben. Das entspricht eine zeitliche Verbesserung von ungefähr 20 Prozent. Auf die Anzahl der korrekten und falschen Antworten vorgreifend, konnten wir auch keine Verschlechterung der Antwortqualität trotz sinkender Bearbeitungszeit feststellen.

Zwischenfazit: CLEF erlaubt eine Bewertung der Anbieter ohne merkbaren (zeitlichen) Zusatzaufwand. Insbesondere wenn man berücksichtigt, dass mit *CLEF* weit mehr untersucht wird als nur die Datenschutzerklärung. Zusätzlich bleibt festzuhalten,

dass Nutzer sich schnell an das System gewöhnen. Insgesamt können wir (F2) positiv beantworten.

Grad der Übereinstimmung zwischen den Teilnehmern Die Übereinstimmung der Antworten unter den Teilnehmern (engl. agreement, concordance, inter-rater reliability) ist ein wesentliches Kriterium für die Bewertung der Qualität unseres Ansatzes. Je höher der Grad der Übereinstimmung zwischen den Teilnehmern, desto weniger kontrovers ist das Verständnis unserer Fragen und die Anwendung der Fragen auf die Anbieter. Wir bewerten die Übereinstimmung anhand zweier unterschiedlicher Maße. Beide Maße berechnen für jeden Anbieter individuell den Grad der Übereinstimmung. u beschreibt den Nutzer, s den Anbieter und q eine Frage.

$$\text{agreement}(\text{u},\text{s},\text{q}) = \frac{\text{\# Antw. auf q identisch zu Antw. von u}}{\text{\# aller Antworten auf q}} \quad [5.1]$$

Erstens berechnen wir (i) für jeden Nutzer die Übereinstimmung seiner Antwort im Verhältnis zu den Antworten aller anderen Teilnehmer, die die gerade betrachtete Frage bei dem gerade betrachteten Anbieter bewertet haben (Gleichung 5.1).

$$\text{agreement}(s) = \frac{\sum_u \sum_q \text{agreement}(\text{u},\text{s},\text{q})}{\text{\# Nutzer die s bewertet haben} * \text{\# Fragen}} \quad [5.2]$$

Daraus berechnen wir anschließend (ii) die Übereinstimmung pro Anbieter, indem wir den Mittelwert aus allen Fragen des betrachteten Anbieters berechnen (Gleichung 5.2).

Zweitens wenden wir Fleiss' Kappa an [JL.71]. Fleiss' Kappa ist ein Maß für den Grad der Übereinstimmung mehrerer bewertender Personen (hier die Teilnehmer) von mehreren verschiedenen Objekten (hier die Fragen). Wie auch in der Vorstudie orientieren wir uns an [LK77]. Kappa-Werte kleiner 0 bedeuten keine Übereinstimmung, zwischen 0 und 0,2 eine geringfügige Übereinstimmung, zwischen 0,2 und 0,4 eine mittelmäßige Übereinstimmung, zwischen 0,4 und 0,6 eine angemessene Übereinstimmung, zwischen 0,6 und 0,8 eine beachtliche und zwischen 0,8 und 1 eine perfekte Übereinstimmung.

Abbildung 5.3 zeigt die kumulative Verteilungsfunktion der Ergebnisse sowohl für den Mittelwert der paarweisen Vergleiche, als auch der Kappa-Werte. Alle Werte sind pro Anbieter berechnet worden. Die prozentuale Anzahl der Anbieter ist auf der vertikalen, die paarweise Übereinstimmung in Prozent und die Kappa-Werte auf der horizontalen Achse abgetragen. Die vertikalen Linien bei 0,4, 0,6 und 0,8 sind die Grenzen für die unterschiedliche Qualität der Kappa-Werte nach [LK77].

Die Auswertung der Graphik zeigt, dass für 20% der Kappa-Werte die Übereinstimmung angemessen ist, das heißt, die Kappa-Werte liegen im linken Drittel des Diagramms. Umgekehrt und für uns vielversprechend deutet für 80% der Anbieter der Kappa-Wert auf eine beachtliche bis zu einer perfekten Übereinstimmung hin.

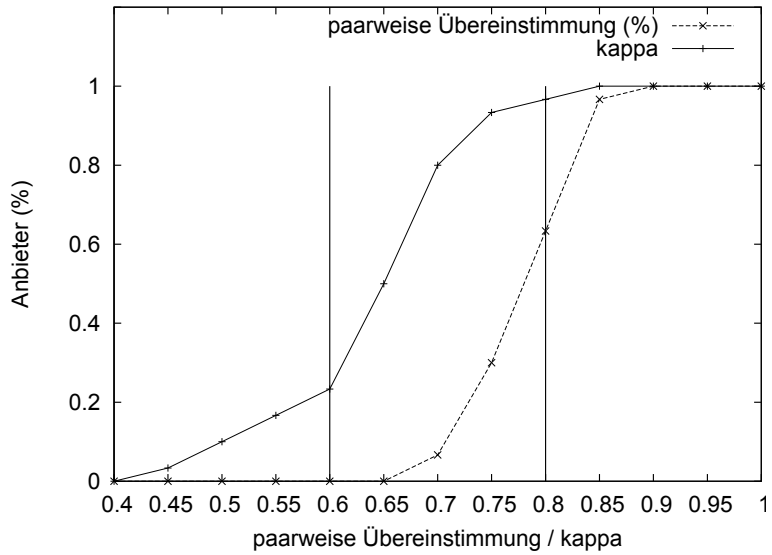


Abbildung 5.3.: Grad der Übereinstimmung unter den Teilnehmern

Die Werte des paarweisen Vergleichs bestätigen diese Aussage und geben eine intuitive Aussage über die Qualität der Mehrheitsentscheidung ab. Für 70% der Anbieter ist die Übereinstimmung der Nutzermeinung größer $\frac{3}{4}$. Abgelesen werden kann dies durch das Legen einer Gerade durch 0,3 auf der vertikalen Achse und dem Schnittpunkt mit der Funktion der paarweisen Übereinstimmung.

Zwischenfazit: Nutzer haben ein sehr hohes Maß an Übereinstimmung bei der Bewertung der Anbieter beziehungsweise bei der Beantwortung der Fragen unserer Taxonomie. Unsere Forschungsfrage (F3) kann somit positiv beantwortet werden.

Richtigkeit der Antworten (Gold-Standard) Wir haben im vorigen Abschnitt gezeigt, dass die Antworten der Nutzer stark übereinstimmen. Der hohe Grad der Übereinstimmung bedeutet jedoch nicht automatisch, dass die Antwort korrekt sein muss, das heißt mit der Meinung der Experten, die die Fragen definiert haben, übereinstimmt. Im Folgenden evaluieren wir, inwieweit die Antworten auch mit unserem Gold-Standard übereinstimmen.

Zuerst untersuchen wir, ob wir die Null Hypothese H_0 = 'zufällige Übereinstimmung der Antworten der Experten und der Teilnehmer' zurückweisen können. Zu diesem Zweck berechnen wir *erstens* für jede Antwort der Nutzer die dichotome Variable {korrekte Antwort, falsche Antwort}. Mit Hilfe eines Binomialtests testen wir dann (i) die Hypothese für alle Nutzer auf einmal und (ii) für die Mehrheitsantwort zu je-

der Frage eines jeden Anbieters. *Zweitens* evaluieren wir für die Mehrheitsantwort auf jede Fragen und für jeden Anbieter die Anzahl (i) der korrekt identifizierten Verstöße (matches:=$V_{gold} \bigcap V_{teilnehmer}$). Der Fehler bei der Identifikation kann zweiseitig sein. Das heißt, wir untersuchen weiter (ii), wie viele Verstöße die Teilnehmer nicht identifiziert haben (misses:=$\frac{V_{gold}}{V_{teilnehmer}}$) und (iii) wie viele Verstöße die Teilnehmer fälschlicherweise erkannt haben (false positives:=$\frac{V_{teilnehmer}}{V_{gold}}$).

Unsere Teilnehmer haben 9.050 (85%) der Fragen korrekt beantwortet. 1.661 (15%) der Antworten stimmen nicht mit dem Gold-Standard überein. Der Binomialtest bestätigt die Zurückweisung von H_0 bei einem Signifikanzniveau von 0,01. Genauso können wir H_0 für die Mehrheitsantwort bei dem gleichen Signifikanzniveau von 0,01 zurückweisen.

Zwischenfazit: Unser Ansatz erlaubt es den Teilnehmern, die Fragen wie von den Experten vorgesehen zu beantworten.

Um darüber hinaus eine Intuition für die Qualität der Antworten der Teilnehmer zu geben, berechnen wir Cohen's Kappa [Coh60]. Cohen's Kappa vergleicht zwei Bewertungen für mehrere Objekte, in unserem Fall also jeden Nutzer einzeln mit dem Gold-Standard. Die bewerteten Objekte sind wie gehabt unsere Fragen. Auch hier kann die Skala von [LK77] angewendet werden.

Wir haben für 79% der Teilnehmer eine beachtliche bis perfekte Übereinstimmung gemessen. Bis auf drei Ausreißer haben alle verbleibenden Teilnehmer immerhin noch eine angemessene Übereinstimmung. Wie beschrieben umfasst unser Gold-Standard 172 Verstöße bei 31 Anbietern. Die Teilnehmer haben gemäß der Mehrheitsmeinung 177 Verstöße identifiziert. 140 Verstöße (81%) wurden von den Teilnehmern richtig erfasst. 19% der Verstöße wurden nicht erkannt, dafür 20% fälschlicherweise.

Zwischenfazit: Unsere Schlussfolgerung aus diesen Ergebnissen ist, dass aufgrund der hohen Übereinstimmung der Antworten verglichen mit unserem Gold-Standard, aber auch aufgrund der starken Übereinstimmung der Meinung der Teilnehmer untereinander, unser Ansatz vielversprechend ist. Wir gehen davon aus, dass Personen ohne Expertenwissen mit *CLEF* viele wesentliche Datenschutzverstöße identifizieren können (F4).

Optimierung der Taxonomie Die Anzahl der nicht erkannten Verstöße (Misses) beziehungsweise der False Positives deutet darauf hin, dass Teile unserer Fragentaxonomie fehlinterpretiert werden können. Wir haben diesen Punkt genauer untersucht, um die Taxonomie zu verbessern. Wir haben herausgefunden, dass drei der 43 Fragen zu überdurchschnittlich vielen Fehlern geführt haben. Ein Beispiel für einen falsch identifizierten Verstoß ist die Bewertung der Experten, dass eine Aussage in der Datenschutzerklä-

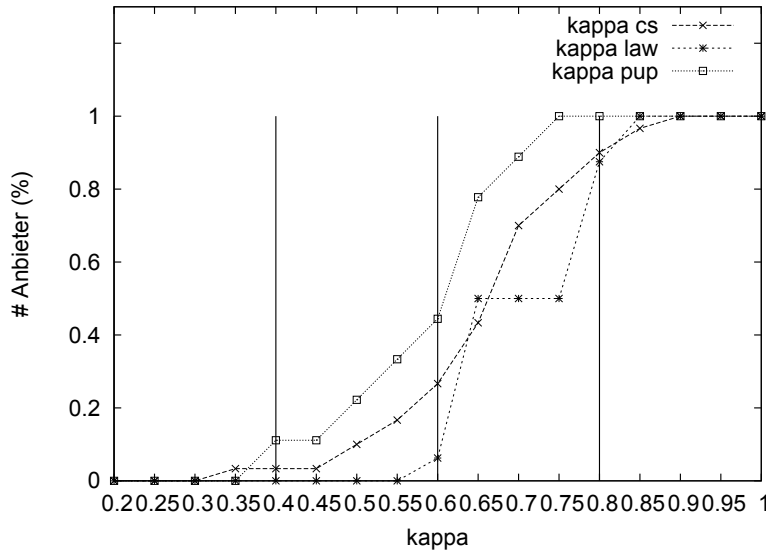


Abbildung 5.4.: Übereinstimmung der Antworten pro Gruppe

rung gemäß 'wir benutzen Cookies für die automatische Anmeldung' ausreichend als Hinweis auf den Einsatz persistenter Cookies ist. Einige Teilnehmer haben dies nicht so bewertet. Betrachten wir die Misses, so haben einige Teilnehmer das Recht, die Einwilligung zu widerrufen, mit dem Recht auf einen Widerspruch (opt-out) verwechselt. Dies ist insbesondere bei dem vom Gesetzgeber geforderten Hinweis auf das Widerspruchsrecht bei der Erstellung pseudonymisierter Profile aufgetreten. Eine Korrektur der Modellierung dieser Verstöße würde zu 88% korrekt identifizierter Verstöße führen, zu 13% Misses und 14% False Positives.

Vergleich der Testgruppen In diesem Abschnitt untersuchen wir, ob die drei Testgruppen der Rechtsstudenten, der Studenten der technischen Wissenschaften und die Schüler unterschiedliche Ergebnisse generiert haben. Wir vergleichen (i) die Zeit, die die Teilnehmer gebraucht haben, um die Anbieter zu bewerten, (ii) das Maß der Übereinstimmung unter den Teilnehmern und (iii) die identifizierten Verstöße.

Wir erwarten, dass die Rechtsstudenten den höchsten Grad der Übereinstimmung der Antworten untereinander aufweisen sowie die meisten Verstöße korrekt identifizieren. Umgekehrt erwarten wir die niedrigsten Werte für die Übereinstimmung und die Erkennung von Verstößen für die Schüler. Trotzdem bewerten wir unsere Ergebnisse als vielversprechend, wenn der Grad der Übereinstimmung und die Anzahl der erkannten Verstöße für alle Gruppen hoch ist.

Tabelle 5.15.: Übereinstimmungen, Misses, False Positives

	Matches	Misses	False Positives
total	0.81	0.19	0.20
cs	0.81	0.19	0.28
law	0.80	0.20	0.20
pup	0.68	0.32	0.14

Vergleich nach Zeit In Bezug auf die Anzahl der Anbieter, die die Teilnehmer bewertet haben, konnten wir nur geringfügige Unterschiede feststellen. Die Rechtsstudenten haben 4,2, die Studenten der technischen Wissenschaften 4,12 und die Schüler 4,38 Anbieter bewertet. Im Mittel haben die Schüler 130 Fragen beantwortet, die beiden anderen Gruppen gleichermaßen jeweils 148 Fragen. Die Schüler haben mehr Zeit für das Lesen der Datenschutzerklärungen und der detaillierten Beispiele, die wir zu jeder Frage beigefügt haben, gebraucht. Trotzdem finden wir den Aufwand für die unterschiedlichen Gruppen vergleichbar. Dass die Schüler mehr Anbieter trotz weniger Fragen beantwortet haben, ist in der größten Anzahl von 'weiß nicht' Antworten begründet. Wir haben im Fall von 'weiß nicht' auf eine abstrakte Frage auch keine weiterführenden Detailfragen gestellt.

Vergleich nach Grad der Übereinstimmung Um den Grad der Übereinstimmung der Antworten der unterschiedlichen Gruppen zu messen, haben wir erneut Fleiss' Kappa angewendet. Die kumulative Verteilungsfunktion der Kappa-Werte zeigt Abbildung 5.4. Jede Kurve repräsentiert eine Gruppe.

Gemäß unserer Erwartung haben die Schüler den niedrigsten Grad der Übereinstimmung. Man sieht dies an der früher ansteigenden Kurve der Schüler im Vergleich zu den anderen beiden Gruppen. Die Gruppe der Rechtsstudenten hat den höchsten Grad der Übereinstimmung. Wir haben jedoch für mehr als 50% der Anbieter eine beachtliche Übereinstimmung gemessen, für die Rechtsstudenten sogar für mehr als 90% der Anbieter. Für die verbleibenden Anbieter besteht immer noch eine angemessene Übereinstimmung (hier nochmal der Hinweis, dass Werte < 0 keine Übereinstimmung bedeuten und die Abbildung erst mit 0,2 beginnt). Wir gehen anhand dieser Ergebnisse davon aus, dass das Verständnis unserer Fragen sehr gut ist. Die Idee unseres Ansatzes, das heißt die Anwendung der Fragen auf die Anbieter, ist für alle Gruppen erfolgversprechend.

Vergleich nach Richtigkeit der Antworten Der Vergleich der Ergebnisse der einzelnen Gruppen bezogen auf die Richtigkeit der abgegebenen Bewertungen zeigt, dass die technischen Studenten und die Rechtsstudenten sich sehr ähnlich verhalten (siehe Tabelle 5.15). Schüler haben hingegen etwa 12% weniger Verstöße identifiziert. Gleich-

zeitig haben die Schüler jedoch auch die niedrigste Anzahl fälschlich identifizierter Verstöße. Unsere Interpretation ist, dass die Schüler aufgrund ihrer geringeren Erfahrung mit Datenschutz, dem Lesen von Verträgen etc. die von *CLEF* ausgewiesenen ausführlichen Beispiele und Beschreibungen intensiver / häufiger zur Beantwortung einer Frage herangezogen haben. Das hat zu einer akkurateren Beantwortung der Fragen geführt. Die geringere Anzahl beantworteter Fragen in der für alle Gruppen gleichen Experimentzeit unterstützt diese These. Die größere Anzahl der Misses ist in der größeren Anzahl an 'weiß nicht' Antworten begründet.

Zwischenfazit: Den Vergleich der unterschiedlichen Gruppen zusammenfassend zeigen unsere Ergebnisse, dass Nutzer ohne jede Vorkenntnis im Kontext Datenschutzrecht eine Vielzahl von Verstößen korrekt identifizieren können. Das Hintergrundwissen über Datenschutz im Speziellen und über Gesetze im Allgemeinen spielt eine untergeordnete Rolle (F5).

5.2.3.3. Diskussion der Hauptstudie

In diesem Abschnitt diskutieren wir unseren Ansatz und geben weitere Argumente, warum unsere Studie aussagekräftig ist.

Angriffe Privatheit ist für viele Unternehmen, wie zum Beispiel für die Betreiber von sozialen Netzwerkseiten, ein Reputationsfaktor. Böswillige Nutzer oder Unternehmen könnten *CLEF* missbrauchen, um beispielsweise Konkurrenten zu schaden. In dieser Arbeit ist es unser Ziel zu zeigen, dass die kollaborative Identifikation von Datenschutzverstößen überhaupt möglich ist. Eine detaillierte Übersicht über Reputationssysteme und die Erkennung von Angriffen ist in [HZNR09, JIB07] zu finden.

Mehrheitsentscheidung Die Mehrheitsentscheidung ist der einfachste Mechanismus, um zu einer Entscheidung zwischen mehreren Teilnehmern zu kommen. In unserer Studie haben wir sie aus zwei Gründen genutzt: *Erstens* wollten wir den Gold-Standard, wie nachfolgend begründet, während der Studie nicht preisgeben. Würde ein Nutzer sehen, dass er viele oder wenige Punkte erhalten hat, könnte er wider besseren Wissens seine Meinung ändern. Die Konsequenz ist allerdings, dass wir auch nicht die Qualität einer Bewertung berechnen können. In der Praxis könnte diese Qualität zum Beispiel anhand der Anzahl der erkannten und durch die Datenschutzaufsicht bestätigten Verstöße berechnet werden. *Zweitens* gilt es zu berücksichtigen, dass die Aspekte bei der Bewertung eines Anbieters vielfältig sind. Wir können nicht davon ausgehen, dass ein Teilnehmer gleichermaßen gute Antworten in jedem Kontext gibt. Beispielsweise bedürfen die Kontexte 'Einsatz von personalisierten Profilen', 'Einsatz automatisierter Verfahren zur Datenverarbeitung' und 'Erfordernis einer Einwilligung' grundsätzlich unterschiedliches Hintergrundwissen. Um also die Antworten verschiedener Nutzer

unterschiedlich zu gewichten, bedarf es unserer Auffassung nach mehr als im Mittel 4 (vergleiche Abschnitt 5.2.3.2) Bewertungen einer Frage.

Perspektive In unserer Studie haben die Teilnehmer die Verstöße von extern identifiziert, also ohne direkten Einblick in das Unternehmen. Eine Erweiterung der Taxonomie würde es aber durchaus ermöglichen, auch Verstöße in Unternehmen von innen heraus zu identifizieren. Die Angestellten würden dann die 'Community' darstellen, gegebenenfalls mit nur Teilwissen spezifisch für ihre Abteilung, und nur die Fragen der Taxonomie beantworten, zu denen sie das Wissen haben. Die Datenschutzverantwortlichen der Unternehmen könnten diese Antworten, die die Praktiken der einzelnen Abteilungen widerspiegeln, mit den Sachverhalten abgleichen, in die der Kunde eingewilligt hat.

5.2.4. Zusammenfassung

Wir haben in diesem Abschnitt einen Ansatz vorgestellt, der es erlaubt, kollaborativ Datenschutzverstöße zu identifizieren. Zu diesem Zwecke haben wir Expertenwissen von Juristen auf eine Taxonomie intuitiver Fragen abgebildet. Diese Fragen werden von Nicht-Experten beantwortet. Zur Erkennung von Verstößen definieren Experten Antwortmuster, die bei Übereinstimmung mit den Antworten der Nicht-Experten auf einen Verstoß hindeuten.

Anhand umfangreicher Nutzerexperimente haben wir gezeigt, dass (F1) Expertenwissen erfolgreich in intuitive Fragen umgesetzt werden kann. Die resultierenden Fragen können in vertretbarer Zeit, das heißt ohne einen wesentlichen Zusatzaufwand im Vergleich zum reinen Lesen und Verstehen der Datenschutzerklärung, durchgeführt werden (F2). Fragen, die gemäß unserer Methodik entwickelt werden, können nachweislich sehr einheitlich beantwortet werden (F3), und Nutzer geben dabei in 81% der Fälle auch die richtige Antwort (F4). Das Hintergrundwissen der Personen spielt eine untergeordnete Rolle (F5).

5.3. Fazit

Die Überprüfung der Einhaltung von Datenschutzgesetzen bei Diensteanbietern im Internet ist schwierig. In einer umfangreichen Analyse des Vollzugsdefizits haben wir gezeigt, dass sich weder die Anbieter konform zu geltenden Gesetzen verhalten, noch die Datenschutzaufsicht die Einhaltung wirksam kontrolliert. So haben die 100 untersuchten Unternehmen zusammen mehr als 300 Verstöße begangen.

Aber nicht nur den Unternehmen fällt es schwer, ihre eigenen Datenschutzverstöße zu identifizieren. Ohne ausreichende Ressourcen gilt dies auch für die Datenschutzaufsichtsbehörden und ohne datenschutzrechtliches Fachwissen auch für den einfachen

Internetnutzer, das heißt den Betroffenen.

Wir haben im zweiten Teil dieser Arbeit einen Ansatz vorgestellt, bei dem Nutzer kollaborativ Datenschutzverstöße identifizieren. Zu diesem Zweck haben wir juristische Expertise in Software gegossen. Die Evaluation anhand einer Nutzerstudie hat gezeigt, dass Nutzer ohne Expertenwissen mit unserem Ansatz große Teile der Verstöße identifizieren, die auch Experten finden.

Wir sehen ausgehend von unseren Ergebnissen mögliche positive, gesellschaftliche Auswirkungen:

Kunden Kunden von Online-Diensten haben neben dem beschriebenen Transparenzproblem der Datenschutzpraktiken der Anbieter das in Kapitel 4 untersuchte Bewusstseinsproblem beim Umgang mit ihren personenbezogenen Daten. Ein Ansatz wie *CLEF* kann sensibilisierend wirken. Die Nutzer von *CLEF* werden auf die relevanten Sachverhalte aufmerksam gemacht und können anhand von Realwelt-Anbietern die Anwendung geltenden Rechts nachvollziehen. Hinzu kommt ein Effekt wie bei Wikipedia, dass in einem kollaborativen System sehr viele Personen von den unter Umständen wenigen Personen, die Beiträge schreiben, profitieren. In unserem Fall sind die Beiträge die Anbieterbewertungen. So könnten die Ergebnisse der Bewertungen als Browser-Plugin beim Besuchen der Seite eines Anbieters angezeigt werde und den Nutzer vor Verstößen warnen.

Unternehmen Vielfach würden Unternehmen gerne alle Datenschutzvorgaben einhalten, sie wissen jedoch nicht wie, beziehungsweise haben keine Möglichkeit, sich selbst zu evaluieren. *CLEF* erscheint vielversprechend für Unternehmen, solch einen Bewertungsmaßstab für eine Selbstüberprüfung darzustellen. Weiter erlaubt es *CLEF* den Unternehmen, Änderungen im Gesetz nachzuvollziehen und nach der Anpassung der eigenen Datenschutzpraktik die Neuerungen wieder selbst auf ihre Korrektheit zu überprüfen.

Datenschutzaufsicht Die Datenschutzaufsicht kann ihre beschränkten Ressourcen auf die Anbieter konzentrieren, bei denen Internetnutzer bereits signifikante Hinweise auf einen Verstoß erkannt haben. Das kommt wiederum den Betroffenen entgegen, da diese darauf vertrauen können, dass unerlaubte Praktiken erkannt und auch verfolgt werden.

Zertifizierungsinstitutionen In unseren Studien ist uns nicht aufgefallen, dass zertifizierte Unternehmen mehr oder weniger Verstöße begehen als nicht zertifizierte. Zertifizierungsinstitutionen würden durch die Anwendung von *CLEF* vergleichbar und könnten einen Verstoßkatalog, wie hier vorgeschlagen, als Mindeststandard vorschreiben.

Zusammenfassend kommen wir zu dem Ergebnis, dass neuartige Ansätze wie *CLEF* vielversprechend sind für die Durchsetzung geltenden Rechts.

6. Anonymisierung als Ausweg für Nutzer und Anbieter

Wir haben bis zu diesem Punkt gezeigt, dass der verantwortungsvolle Umgang mit personenbezogenen Daten sowohl die Unternehmen als auch die von der Verarbeitung personenbezogener Daten betroffenen Personen überfordert. Außerdem bestehen zwei unterschiedliche Arten von Schutzbedürfnissen: der Schutz zwischen Nutzern und der Schutz von Nutzern gegenüber Anbietern.

In diesem dritten und letzten Teil der Arbeit führen wir diese Aspekte am Beispiel der Protokollierung von Suchanfragen zusammen. Das heißt, wir zeigen einen Ausweg auf, so dass die Privatheit der Nutzer gewahrt bleibt, dass Anbieter von vielen Anforderungen durch den Gesetzgeber entbunden werden und dass die Anbieter gleichzeitig und weiterhin Nutzen aus den protokollierten Daten ziehen können. Dazu greifen wir drei bisherige Ergebnisse wieder auf: (i) Die Studie zu der kollaborativen Suchmaschine (Abschnitt 4.1) hat gezeigt, dass Nutzer für den gegenseitigen Austausch von Suchanfragen die Möglichkeit zur Anonymität fordern. (ii) Die größte Sorge der CSE-Nutzer ist, dass die Anbieter Profile für beispielsweise personalisierte Werbung über sie erstellen. Nun ist es gerade die personalisierte Werbung, die die Haupteinnahmequelle für Suchmaschinenanbieter darstellt, das heißt den Nutzen. Und gerade ihretwegen können Anbieter die Suche im Internet und viele Zusatzdienste kostenlos anbieten. (iii) Rückmeldungen auf unsere Vollzugsdefizitanalyse (Kapitel 5.1) zeigen, dass viele Unternehmen versuchen, konform zu geltendem Recht zu sein, jedoch mangels Kontrollmöglichkeiten, beispielsweise wie durch *CLEF* (Kapitel 5.2), scheitern.

6.1. Privatheitsprobleme durch Suchhistorien

[WLW98] fordert eine Ausgewogenheit zwischen Privatheit und den Vorteilen eines uneingeschränkten Informationsflusses. In diesem Kapitel stellen wir einen auf Anonymisierung basierenden Lösungsansatz vor.

Nutzerperspektive Die (versehentliche) öffentliche Preisgabe von Suchprotokollen durch AOL im Jahr 2006 [PCT06] hat gezeigt, dass Suchprotokolle die eindeutige Identifizierung von Nutzern erlauben [BTZ06, Kan06]. 2008 hat sich die Europäische Gemeinschaft dieser Auffassung angeschlossen [Dat08]. Mit anderen Worten, das Speichern und Verarbeiten von Suchprotokollen gefährdet die Privatheit der Nutzer. Auch

die Teilnehmer unserer CSE-Studie fühlen sich durch umfangreiche Suchprotokolle bedroht – sowohl gegenüber anderen Nutzern als auch gegenüber dem Anbieter.

Anbieterperspektive Werbung ist die Haupteinnahmequelle von Suchmaschinenanbietern (im Folgenden als Anbieter bezeichnet). Anbieter versuchen zum optimalen Auswählen und Positionieren von Werbung die Intention des Nutzers, das heißt sein aktuelles Informationsbedürfnis, vorherzusagen. Der Umsatz der Unternehmen hängt von der Vorhersagequalität ab. Um die Positionierung von Werbung zu verbessern, planen die Anbieter den Einsatz umfangreicher Suchprotokolle. Ihre Datenschutzerklärungen zeigen, dass sie sowohl in der Lage als auch gewillt sind, dies zu tun[1][2][3].

Außerdem müssen Anbieter bei der Speicherung und Verarbeitung personenbezogener Daten eine Vielzahl datenschutzrechtlicher Anforderungen erfüllen. Meistens ist diese Erfüllung schwierig und nicht zuletzt kostenintensiv. Die Auflagen umfassen beispielsweise für bestimmte Praktiken das Einholen einer Einwilligung vom Nutzer, Informationspflichten, das Durchlaufen oftmals schwieriger Prozesse zum Löschen von Daten, etc. [Par95, Dat08]. In Fällen, in denen Anbieter diese Auflagen nicht korrekt erfüllt haben und der Verstoß publik wurde, hat das oftmals eine deutlich negative Presse nach sich gezogen [BTZ06].

Eine Möglichkeit, sowohl den Forderungen der CSE-Nutzer als auch denen der Anbieter gerecht zu werden, ist Anonymisierung. Der Einsatz anonymisierter Suchprotokolle würde sowohl die Privatheit der Nutzer verbessern als auch die Anbieter von den gesetzlichen Pflichten entbinden (siehe auch Abschnitt 2.4.2). Gemäß [Par95] ist ein Datensatz anonym, wenn keine Person aus dem Datensatz persönlich identifiziert werden kann.

[TMK08] hat Anonymität für mengenwertige Daten definiert. Wir verweisen darauf als (k,m)-Anonymität. Nach der Definition muss jede Kombination von m Termen aus der Suchhistorie eines Nutzers ebenfalls in den Suchhistorien von $k-1$ anderen Nutzern vorkommen. Die Instanz, die die Daten anonymisiert, entscheidet sich dabei für den Parameter m, also die Anzahl der Terme, die einen potentiellen Quasi-Identifikator formen können. Wir werden diese Definition von Privatheit in diesem Kapitel verwenden.

Beispiel 21: [Swe00, Gol06] haben gezeigt, dass ungefähr 67% der US Bürger durch {Geburtsdatum, Postleitzahl, Geschlecht} eindeutig identifiziert werden können. $m = 3$ bei der Anonymisierung der Daten schützt vor solch einer Re-Identifikation, da (k,3)-Anonymität garantieren würde, dass jede Kombination von drei Attributen bei mindestens k Nutzern existiert.

[1] `http://www.google.co.uk/intl/en/privacy_cookies.html`, Nov. 2009
[2] `http://privacy.microsoft.com`, Januar 2010
[3] `http://privacy.yahoo.com`, Januar 2010

Je höher *m*, desto größer die Anzahl Attribute, die ein Angreifer kennen muss, um ein Individuum zu identifizieren. Je höher *k*, desto mehr Nutzer sind voneinander ununterscheidbar.

Wir halten diese Definition von Anonymität für unser Szenario aus den folgenden Gründen für nützlich: Anonymitätsdefinitionen unterscheiden sich in den Annahmen darüber, was privat sein soll. So kann man nicht einfach Definitionen von Anonymität für Krankenakten übernehmen [Swe02, MGKV06, LLV]. Bei Krankenakten weiß die die Daten anonymisierende Instanz, welche Daten eine Person identifizieren, zum Beispiel die Attribute Name und Versicherungsnummer. Ebenso weiß sie, welche Daten sensibel sind, zum Beispiel die Krankheit oder die Medikamentierung. Bei Suchhistorien gibt es dieses Vorwissen, welche Terme oder Termkombinationen zu einer Re-Identifikation führen können oder welche sensibel sind, nicht.

Gegeben eine Definition von Anonymität, so gibt es unterschiedliche Wege, ein Suchprotokoll so zu verändern, dass es der Definition entspricht. Man kann Daten löschen, Daten hinzufügen, Daten zwischen Datensätzen austauschen oder die Daten modifizieren, das heißt generalisieren. Mit Techniken, die Daten zum Zwecke der Anonymisierung einfügen [MC08b], wird die Vorhersage der Intentionen der Nutzer deutlich schwieriger – sie wirken den Interessen der Anbieter, Werbung möglichst intelligent zu schalten, klar entgegen. Techniken, die hingegen eine vollständige Generalisierungshierarchie fordern [TMK08, HN09], versagen, wie in Kapitel 2.4.2 beschrieben. Das liegt daran, dass das Bilden einer solchen Hierarchie eine fast unlösbare, weil niemals endende Aufgabe darstellt. Das heißt wiederum, dass die Anonymisierung von Suchhistorien basierend auf Generalisierung nur für sehr beschränkte Szenarien möglich ist. Eigenschaften solcher Szenarien umfassen ein konstantes Vokabular, beschränkte Dimensionalität des Vokabulars oder den Ausschluss von Homonymen. Um (k,m)-Anonymität zu erreichen, werden wir Terme aus den individuellen Historien der Nutzer solange löschen, bis die gewünschte Anonymität erreicht ist.

Mit Anonymisierung geht eine Abwägung zwischen Privatheit und dem Nutzen der anonymisierten Daten einher [Ada07]. In unserem Fall bedeutet das aus Sicht der (CSE-) Nutzer die verbleibende Möglichkeit, Anfragen und Links auszutauschen. Aus Sicht des Anbieters ist der Nutzen insbesondere bestimmt durch die Verwendbarkeit der anonymisierten Daten zu Werbezwecken. Es gibt für ein Suchprotokoll mehrere Wege, den gleichen Grad Privatheit zu erreichen, bei einem unterschiedlichen Nutzen des resultierenden Datenbestandes.

Existierende Arbeiten [Ada07, TMK08, HN09] schlagen Anonymisierungsalgorithmen vor und untersuchen deren Effektivität. Jeder der Algorithmen zielt auf ein einziges Merkmal der anonymisierten Daten ab, zum Beispiel versuchen sie die Anzahl erforderlicher Generalisierungen kleinzuhalten oder die Größe des anonymisierten Suchprotokolls zu maximieren. Solch eine Betrachtung ist jedoch bis zu einem gewissen Grad undifferenziert: Beispielsweise könnte, bei gleicher Anonymität, die ‘optimal’ anonymisierte Historie für unterschiedliche Geschäftsmodelle der Anbieter, wie Be-

zahlung pro Werbeeinblendung (engl. Pay-Per-Impression), pro Klick auf eine Werbung (engl. Pay-Per-Click) oder auch nach dem durch die Werbung generierten Umsatz (engl. Pay-Per-Revenue), variieren. Alternativ, da die Werber ebenfalls unterschiedliche Geschäftsmodelle haben können, kann 'optimal' auch das Bieten auf spezifische, teure Terme oder unspezifische, günstige Terme bedeuten. Die CSE-Nutzer möchten unter Umständen möglichst viele Terme austauschen können und das mit möglichst vielen Personen. Diese vielen verschiedenen Anforderungen machen deutlich, dass Algorithmen zur Anonymisierung bezüglich der Zielfunktion flexibel sein müssen. Entsprechend erfordert es für die Aussage, was eine 'optimale' Anonymisierung ist, eine systematische Untersuchung der möglichen Zielfunktionen.

Beiträge dieses Kapitels Basierend auf der Definition von (k,m)-Anonymität evaluieren wir den Einfluss unterschiedlicher Zielfunktionen [1]. Wir tun dies zum einen mit der Perspektive auf den Nutzer, das heißt in Hinblick auf Nutzung anonymisierter Suchprotokolle bei CSEs. Zum Anderen bewerten wir den Einfluss der Zielfunktionen mit der Perspektive auf den Anbieter, das heißt auf die Möglichkeit, Werbung zu schalten. Wir berücksichtigen als Zielfunktionen generische Charakteristiken wie 'Größe des anonymisierten Suchprotokolls' und die 'Anzahl Nutzer', für die eine Anonymisierung der Suchhistorie möglich ist. Diese sind interessant im Hinblick auf die CSE-Nutzer. Speziell für den Aspekt Werbung betrachten wir Zielfunktionen, die auf Terme mit einem hohen Marketingwert abzielen, auf Terme mit vielen Werbeeinblendungen, etc. Unter Berücksichtigung dieser Charakteristiken untersuchen wir, wie mit verschiedenen Zielfunktionen anonymisierte Suchprotokolle geartet sind. Wir tun dies darüber hinaus für variierende Grade von Privatheit.

Zu diesem Zweck haben wir eine Heuristik entwickelt, die (k,m)-Anonymität für Suchhistorien schafft. Unsere Implementierung ist flexibel bezüglich der Zielfunktion.

Die Realweltsuchhistorien, die wir für unsere Evaluierung einsetzen, umfassen 3,5 Millionen Anfragen von 370.585 Nutzern. Um ebenfalls Realweltmarketingdaten zu erhalten, haben wir das Yahoo! Marketingportal systematisch für jeden Term aus den Suchhistorien aller Nutzer ausgelesen. Dieses Portal bietet Werbern Informationen speziell zur Planung von Werbekampagnen, beispielsweise wie oft ihre Werbung voraussichtlich angeklickt wird, wie oft sie angezeigt wird etc.

Unsere Hauptergebnisse in diesem Kapitel sind, dass (k,m)-Anonymität vor der Re-Identifikation ähnlich derer im AOL-Fall schützt. Sie bewahrt Inhalte, die CSE-Nutzern beim Suchen helfen und Suchmaschinenanbietern wie Werbern bei der Verfolgung ihrer ökonomischen Interessen – mit einer deutlich reduzierten Gefahr für die Privatheit und reduzierten rechtlichen Anforderungen an den Anbieter. Aus Sicht der CSE-Nutzer sind zwischen 18% und 45% aller Suchterme auch anonymisiert bei kollaborativen Suchmaschinen nutzbar, abhängig vom Grad der Privatheit. Eine Anonymisierung der Suchhistorien ist für 70% bis 95% aller Nutzer möglich. Aus Sicht der Anbieter wahrt

eine Anonymisierung für m=3, k=100 die Terme, die nach unserer Abschätzung immerhin noch zu 61% der Klicks auf Werbeanzeigen führen würden. Außerdem zeigen wir, wie wichtig die Wahl der Zielfunktion bei der Anonymisierung ist. So kann bei dem gleichen Grad der Anonymität (zum Beispiel m=3, k=100), die Wahl der falschen Zielfunktion zu 40% weniger Klicks auf Werbung führen.

Der weitere Aufbau dieses Kapitels ist wie folgt: In Kapitel 6.2 geben wir Hintergrundinformationen zu Werbung bei Suchmaschinen und (k,m)-Anonymität. In Abschnitt 6.3 stellen wir unseren Ansatz vor und evaluieren diesen in Abschnitt 6.4. Abschnitt 6.5 fasst unsere Ergebnisse zusammen.

6.2. Hintergrundinformationen

In diesem Abschnitt betrachten wir, welche Daten Suchmaschinenanbieter aktuell von ihren Nutzern erheben (Abschnitt 6.2.1) und den Status Quo von Werbung bei Suchmaschinen (Abschnitt 6.2.2). Darüber hinaus beschreiben wir die Herausforderungen bei der Anonymisierung von Suchhistorien (Abschnitt 6.2.3). Zuletzt führen wir das Konzept der Anonymisierung mengenwertiger Daten ein und definieren (k,m)-Anonymität (Abschnitt 6.2.4).

6.2.1. Datenerhebung von Suchmaschinenanbietern

Anbieter erheben aktuell große Mengen personenbezogener Daten. Eine wichtige Motivation dahinter ist personalisierte Werbung. Microsoft, Google und Yahoo! geben genau diesen Grund in ihren relevanten Datenschutzerklärungen an[1][2][3]. Ein Anbieter erhebt Daten von Nutzern, die sich registrieren, aber auch von nicht registrierten Nutzern. In beiden Fällen beinhaltet dies Cookie-Informationen, die IP-Adresse, – und für uns relevant – Suchanfragen. Während die Anonymisierung von IP-Adressen relativ auf der Hand liegend ist, zum Beispiel durch Ausmaskieren der niederwertigen Bits der Adresse, ist dies bei Suchanfragen deutlich schwieriger. Wenn auch mit der Einwilligung des Nutzers, erheben die Anbieter außerdem den Namen, die Postleitzahl, das Geschlecht und das Geburtsdatum von Personen mit beispielsweise einer Windows-Live-ID oder einem Google-Zugang. Logischerweise können diese Angaben mit den Suchanfragen korreliert werden. In dieser Arbeit schließen wir solche Verknüpfungen von zusätzlichem Wissen jedoch aus, da dies eine eigene Forschungsfrage darstellt.

Für die registrierten Nutzer bieten die Anbieter die Möglichkeit, ihre Einwilligung zu widerrufen oder bestimmten Praktiken zu widersprechen – für die Nutzer, die die Datenschutzerklärung finden, lesen und verstehen. In vielen Fällen entspricht das zumindest grob auch dem, was der Gesetzgeber fordert. In Kapitel 5.1 haben wir jedoch gezeigt, dass in der Praxis Mechanismen wie das Widerrufen einer Einwilligung oder das Löschen von Daten auch bei Suchmaschinen nicht funktionieren.

Nun kann man annehmen, dass nicht registrierte Nutzer einen höheren Grad von Privatheit haben. Unglücklicherweise ist dies nicht der Fall. Lange Suchhistorien an sich können ein effektives Nutzerprofil darstellen [JKPT07] und wirken häufig identifizierend [BTZ06]. Google erstellt Suchhistorien auch für unregistrierte Nutzer[4]. Hat der Nutzer diese Funktion bewusst deaktiviert, führt ein Browserwechsel oder der Wechsel des PCs dazu, dass Historien wieder gebildet werden, da diese Funktion standardgemäß aktiviert ist.

Anbieter wiederum sind konfrontiert mit verschiedensten Verpflichtungen resultierend aus der Datenschutzgesetzgebung. Besonders komplex gestaltet sich das für den Anbieter, wenn dieser, wie bei Suchmaschinen üblich, international tätig ist. Die Verarbeitung von nur anonymisierten Daten würde gesetzliche Anforderungen, zum Beispiel bei Auskunftsersuchen, deutlich reduzieren. So müsste ein Anbieter, der ausschließlich anonymisierte Daten verarbeitet, keine Mechanismen zum Widerrufen der Einwilligung zu personalisierten Profilen, dem Löschen von Daten, Speicherlimitierungen, etc. anbieten.

6.2.2. Werbung bei Suchmaschinen

Zur Zeit der Entstehung von Suchmaschinen haben die Anbieter ihren Werbekunden Kosten pro Einblendung einer Anzeige berechnet. Heute zahlen Werber zumeist pro Klick auf ihre Anzeige. Folglich ist aus Sicht des Anbieters der Schlüssel zu einem erfolgreichen Werben das Schalten genau der Anzeigen, die im Interesse des jeweiligen Nutzers sind. Um dies zu erreichen, bieten die Anbieter Plattformen an, mit deren Hilfe Werber in Auktionen auf Schlagworte (engl. ad-words) bieten können. Eine solche Plattform ist beispielsweise Google AdWords[5]. Der dabei am weitesten verbreitete Auktionsmechanismus ist die generalisierte Zweitpreisauktion [EOS07, Var07]. Der Werber, der die Auktion gewinnt, bekommt den besten Anzeigenplatz auf einer Seite, der zweite den zweitbesten Platz. Das geht so weiter, bis keine Werbefläche mehr zu versteigern ist. Sucht ein Nutzer nun nach dem Schlagwort, auf das ein Werber geboten hat, werden die Anzeigen der Gewinner der Auktion eingeblendet. Mittlerweile nutzen die Anbieter bei der Auswahl der Werber auch noch weitere Kriterien. So prüfen sie unter anderem, ob der Text in der Werbeanzeige auch zu der Webseite passt, auf die die Anzeige verweist.

Die Bezahlung pro Klick auf eine Werbeanzeige ist für den Werber deutlich günstiger. Der Nutzer wird vom Anbieter gezielt auf die Seite des Werbers hingewiesen und es entsteht für den Anbieter ein Anreiz, dies, zur Maximierung des Umsatzes, besonders gut zu machen. Wie das folgende Beispiel zeigt, ist dieser Schritt jedoch schwierig:

Beispiel 22: Die Anfrage 'Golf London' auf `www.google.co.uk` liefert fünf Werbeanzei-

[4] `http://www.google.com/support/accounts/bin/answer.py?answer=54081`, Dezember 2009
[5] `https://adwords.google.com`, Juli 2010

gen für Golfkurse, eine für Golfurlaube und zwei ‘generische’ Anzeigen. Startet man nun die Sequenz von Anfragen ‘compact cars’ und ‘corolla golf focus comparison’, mit einigen Klicks auf URLs, die ein Interesse an Autos erkennen lassen, liefert die erneute Anfrage ‘Golf London’ – die gleichen Golfkursanzeigen.

Bei der Vorhersage der Nutzerintention besteht offentlichtlich noch Verbesserungspotential. Die Erwartung, verbunden mit der Speicherung von Suchhistorien, ist, dass die Auswertung (a) aufeinanderfolgender Anfragen zu einem effektiveren Platzieren von Anzeigen führt. Außerdem wird angenommen (b), dass, wenn Anfragen keine beworbenen Schlagworte beinhalten, es längere Suchhistorien dem Anbieter erlauben, solche Anzeigen zu schalten, die zumindest im generellen Interesse der suchenden Person liegen.

6.2.3. Herausforderungen bei der Anonymisierung von Suchhistorien

Das Anonymisieren von Suchhistorien ist schwierig: *Erstens* gibt es bei der Anonymisierung eine inhärente Abwägung (Tradeoff) zwischen der Privatheit und der Qualität der aus der Anonymisierung resultierenden Daten [Ada07]. Hier bedeutet Qualität zweierlei: Aus Sicht des CSE-Nutzers zum Beispiel die Anzahl der Terme, die die Nutzer anonymisiert austauschen können. Aus Sicht des Suchmaschinenanbieters bedeutet es die Möglichkeit, basierend auf anonymisierten Daten Werbung zu schalten. Außerdem gibt es unterschiedliche Geschäftsmodelle der Anbieter, wie ‘Pay-Per-Click’ oder ‘Pay-Per-Impression’. Ebenso gibt es unterschiedliche Strategien der Werber, zum Beispiel ‘viele Werbeeinblendungen aber wenige Klicks’, ‘wenige Einblendungen aber eine hohe Klickrate’ etc. [EOS07]. Folglich tendieren Anbieter und Werber zu unterschiedlichem Verhalten, das heißt, sie fordern verschiedene Zielfunktionen. Der Anbieter ist an einem hohen Umsatz interessiert, während der Werber viele Werbeeinblendungen oder Klicks erzielen möchte. Es ist unklar, welche Zielfunktion eine die Daten anonymisierende Instanz einsetzen soll.

Zweitens hat die die Anonymisierung durchführende Instanz kein a priori Wissen darüber, welche Suchterme oder Termkombinationen sensibel sind oder welche zu der Identifikation eines Nutzers führen können. Folglich sind allgemein bekannte Anonymisierungstechniken für Gesundheitsdaten nicht anwendbar.

Drittens ist der Einsatz von Generalisierungshierarchien wie in [HN09, TMK08], für Suchhistorien nur schlecht möglich. Das kommt daher, dass solche Hierarchien für den Suchmaschinenkontext mit hunderttausenden von unterschiedlichen Termen nicht existieren. Existierende Hierarchien wie WordNet [Mil95] funktionieren schlecht. Beispielsweise führt die Verknüpfung von WordNet mit dem Altavista-Suchprotokoll zum Löschen von 85% aller unterschiedlichen Terme aus dem Suchprotokoll.

Viertens scheint das Einfügen von Anfragen in das Suchprotokoll zur Verschleierung der Nutzerintention eine Alternative zu sein [MC08b]. Tatsächlich schützt das vor der Re-Identifikation des Nutzers. Es macht es dem Anbieter aber nahezu unmöglich, Wer-

Tabelle 6.1.: Re-Identifikation von Nutzer 4417749

Term	m	Häufigkeit
Georgia	1	1380
Arnold	1	167
Lilburn	1	6
Arnold & Georgia	2	15
Lilburn & Georgia	2	5
Arnold & Lilburn	2	1
Arnold & Lilburn & Georgia	3	1

bung personalisiert zu schalten – den Kompromiss zwischen Privatheit und Nutzen, den wir suchen.

Fünftens ist das Finden einer optimalen Anonymisierung (basierend auf k-Anonymität) NP-schwer [MW04]. Das gilt ebenso für (k,m)-Anonymität, da für die Anzahl der identifizierenden Attribute $m = 1$, (k,m)-Anonymität identisch zu k-Anonymität ist.

6.2.4. (k,m)-Anonymität

In diesem Abschnitt beschreiben wir zuerst, warum mengenwertige Anonymisierung, das heißt (k,m)-Anonymität, zu unserem Szenario passt. Anschließend zeigen wir, wie man Personen in nicht anonymisierten Suchprotokollen identifizieren kann. Wir führen die Definition von (k,m)-Anonymität ein und stellen den Bezug zu der versehentlichen Preisgabe von Suchprotokollen durch AOL her.

Die Entscheidung für (k,m)-Anonymität anstelle irgendeiner der Anonymisierungen, die in [MC08b, Ada07, KNPT07, KKMN09] vorgeschlagen werden, haben wir aufgrund der folgenden Beobachtungen und Annahmen getroffen: Man kann bei der Betrachtung von Suchprotokollen feststellen, dass die Kombination von Termen aus unterschiedlichen Anfragen die Privatheit des Nutzers gefährdet, nicht nur Terme aus einer Anfrage. Beispiel 23 wird das näher illustrieren. Außerdem nehmen wir an, dass einmal preisgegebene Terme ebenso einen Quasi-Identifikator bilden können, wie es auch solche Terme tun können, die ein Nutzer mehrfach verwendet. Beispielsweise gefährdet die einmalige Preisgabe der Kreditkartennummer, die in vielen Suchprotokollen zu finden ist, die Privatheit eines Nutzers genauso wie sie es bei einer mehrmaligen Preisgabe tun würde. Mit anderen Worten, die Privatheitsbedrohung ist unabhängig von ihrer Häufigkeit im Suchprotokoll. Weiter können wir nicht antizipieren, welche Kombination von Termen zu der Identifikation eines Nutzers führen wird oder welche Terme sensible Information beinhalten. Entsprechend behandelt die Definition von (k,m)-Anonymität alle Terme gleich, das heißt, sie garantiert Anonymität für jede mögliche Termkombination.

Um die eben beschriebenen Sachverhalte zu illustrieren, schauen wir uns jetzt die versehentliche Preisgabe von Suchhistorien durch AOL genauer an [PCT06].

Beispiel 23: Im Jahr 2006 hat AOL Millionen von Anfragen mit 454 Anfragen und 356 unterschiedlichen Termen von Nutzer 4417749 preisgegeben. Die Historie für diesen Nutzer beinhaltete Suchen nach Personen namens 'Arnold' ebenso wie nach Shops und Landschaftsgärtner in 'Lilburn', 'Georgia'. 166 Personen außer Nutzer 4417749 hatten ebenfalls nach 'Arnold', 5 nach 'Lilburn' und 1379 andere nach 'Georgia' gesucht (Tabelle 6.1, $m = 1$). Folglich ist dieser Datensatz k-Anonym für k=6. Gleichzeitig kann man jedoch der Tabelle entnehmen, dass 'Arnold & Lilburn' einen Quasi-Identifikator bilden, das heißt eine Person identifizieren. Das (und ein Telefonbuch) erlaubte es Reportern herauszufinden, dass 'Thelma Arnold', Wohnhaft 'Lilburn', 'Georgia' Nutzer 4417749 ist.

Wie das Beispiel impliziert, müssen wir eine Definition von Anonymität finden, die Kombinationen von Termen berücksichtigt. [TMK08] hat die folgende Definition für das Warenkorb-Szenario definiert. Wir haben sie für den Kontext Suchprotokolle leicht angepasst:

Definition 4 (k,m)***-Anonymität*** *Ein Suchprotokoll ist (k,m)-anonym, wenn jede Kombination von m Termen aus der Suchhistorie eines Nutzers in wenigstens $k-1$ Historien anderer Nutzer ebenfalls existiert. Weiter nennen wir (k,m) den Grad der Privatheit des anonymisierten Suchprotokolls.*

Betrachtet man nochmal die Häufigkeit der Termkombinationen in Beispiel 23, so sieht man, dass 'Arnold & Lilburn' einen Quasi-Identifikator für Kombinationen der Länge $m = 2$ bilden. Das heißt, das AOL-Suchprotokoll ist für $m > 1$ nicht anonym. (k,m)-Anonymität hätte sichergestellt, dass auch diese Termkombination mindestens k-mal vorkommt. Mit $k = 2$, $m = 2$ wäre Thelma Arnold nicht eineindeutig identifiziert worden.

Merke, dass k und m unabhängig sind. Ist ein Suchprotokoll jedoch anonym bezüglich (k,m), so ist es das auch für (k-1,m). Das gleiche gilt für den Parameter m, das heißt, ein Protokoll, das (k,m)-anonym ist, ist auch (k,m-1)-anonym.

6.3. Ansatz

Wir identifizieren zuerst relevante Zielfunktionen für CSE-Nutzer und den Kontext Werbung bei Suchmaschinen (Abschnitt 6.3.1). Eine genaue Problembeschreibung gibt (Abschnitt 6.3.2). Danach (Abschnitt 6.3.3) beschreiben wir die Umsetzung unseres Ansatzes zur Beantwortung der Fragestellung aus der Problembeschreibung.

Tabelle 6.2.: Illustration unterschiedlicher Nutzenkriterien

Term	# Terme	# Nutzer	max Gebot ($)
or	**142041**	2170	0.42
com	74453	**31974**	0.69
MacMall	4	4	**919**
ATT	270	146	0.5
Lowes	173	101	0.05
AdultFriendFinder	6	3	334
Term	# Einblendungen	# Klicks	Umsatz ($)
or	56574	28	12
com	203681	549	380
MacMall	5775	786	722334
ATT	**5009356**	3747	1874
Lowes	491515	**211489**	10575
AdultFriendFinder	10020	2742	**915828**

6.3.1. Informationsverlust und Nutzen

Wir beschreiben jetzt die unterschiedlichen Zielfunktionen. Die werbespezifischen Zielfunktionen haben wir aus der Marketingplattform von Yahoo! abgeleitet. Dabei handelt es sich um Informationen, die Yahoo! seinen Werbern anbietet, damit diese ihre Kampagnen planen und optimieren können.

 # unterschiedlicher Terme CSE-Nutzer profitieren von einer Vielfalt von Termen. Diese ist gleichbedeutend mit einer hohen Wahrscheinlichkeit, ähnliche Suchanfragen von der CSE angeboten zu bekommen, die bei dem Finden von Informationen helfen. Anbieter könnten ebenfalls gewillt sein, eine maximale Anzahl unterschiedlicher Terme im anonymisierten Suchprotokoll zu wahren. Das kann beim Verstehen der Nutzerintentionen hilfreich sein. Außerdem erlaubt dies das Platzieren thematisch breit gestreuter Anzeigen.

 # Terme Je größer ein Suchprotokoll ist, desto eher können CSEs häufige und damit relevante Suchanfragen und Ergebnisse identifizieren und den Nutzern zur Suchunterstützung anbieten. Anbieter können ebenfalls versucht sein, ein Suchprotokoll maximaler Größe zu erreichen, zum Beispiel, wenn häufige Vorkommen eines Terms wichtiger sind als unterschiedliche, aber seltene Terme. Das könnte dann relevant sein, wenn mehrere Werber auf den gleichen Term bieten möchten, um ihre Werbung angezeigt zu bekommen.

 # Nutzer Das Löschen von Termen aus den Historien der Nutzer kann zur vollständigen Löschung eines Nutzers führen. CSEs sind dann effektiv, wenn viele Nutzer mit unterschiedlichen Interessen das System nutzen. Ziel aus Sicht der CSE

kann es also sein, möglichst viele Nutzer nach der Anonymisierung im Suchprotokoll zu haben. Anbietern kommt eine große Anzahl Historien unterschiedlicher Nutzer entgegen, wenn sie durch die darin enthaltenen Informationen zu diesen Nutzern Werbung schalten können, auch wenn die Information unter Umständen vage ist.

max. Gebot Die meisten Anbieter werden von den Werbern pro Klick auf eine Werbeeinblendung bezahlt. Folglich möchte sie vielleicht gerade die Terme behalten, auf die Werber viel bieten. Wir berechnen durch $\mathsf{bid} \cdot \mathsf{term\_frequency}$ den Gesamtnutzen, das heißt, wir gewichten das Gebot mit der Häufigkeit des Terms. Wir machen das analog bei allen weiteren Zielfunktionen.

 # Klicks Abhängig vom Geschäftsmodell (pay-per click) könnten Anbieter solche Terms halten wollen, zu denen Werbung häufig angeklickt wird, im Gegensatz zu Termen mit hohen Geboten aber wenigen Klicks.

 # Werbeeinblendungen Ein alternatives Geschäftsmodell ist die Bezahlung pro Werbeeinblendung (engl. ad impression). Werbeeinblendung gibt dabei an, wie oft eine bestimmte Werbung angezeigt wird, unabhängig von der Anzahl Klicks.

Umsatz Dieses Kriterium zielt darauf ab, solche Terme im Protokoll zu behalten, bei denen $sum_{term}\mathsf{clicks(term)} \cdot \mathsf{bid(term)}$ maximal ist.

Beispiel 24: Für das Suchprotokoll, das wir in der Evaluierung nutzen werden, gibt Tabelle 6.2 an, welcher Term am häufigsten ist, welchen Term die meisten Nutzer mindestens einmal benutzt haben, den Term mit dem höchsten empfohlenen Gebot und den Term, für den die erwartete Anzahl Einblendungen, Klicks und der Umsatz maximal ist. Offensichtlich führen Sortierungen gemäß der unterschiedlichen Kriterien auch zu unterschiedlichen Positionen der Terme in der Rangfolge. Entsprechend ist davon auszugehen, dass unterschiedliche Zielfunktionen auch das Anonymisierungsergebnis beeinflussen.

Tatsächlich stellen ‘Gebot’ und ‘Umsatz’, so wie wir sie eingeführt haben, nicht den wirklichen Umsatz eines Anbieters dar. Das liegt daran, dass die Anbieter meistens eine generalisierte und gegebenenfalls modifizierte Zweitpreisauktion anbieten, mehrere verschiedene Arten von Werbung gleichzeitig einblenden etc. Nach unserem Wissen ist dies jedoch die erste Studie, die den Nutzen von Anonymisierung anhand von Realwelt-Marketingdaten für Werbezwecke misst.

6.3.2. Problembeschreibung

Es ist eine offene Frage, ob anonymisierte Suchprotokolle einen Wert für Marketingzwecke und das Platzieren von Werbung haben. Um ein (k,m)-anonymes Protokoll S' aus dem initialen Protokoll S zu erhalten, löschen wir, wie wir beschreiben werden, Terme aus den Suchhistorien der Nutzer. Das Protokoll S ist eine Menge von Suchhistorien unterschiedlicher Nutzer. In diesem Kapitel ist eine Historie eine Menge von Termen, die ein Nutzer in seinen Suchanfragen verwendet hat. Das Kritische bei der Anony-

misierung ist die Wahl des Terms, der aus der Historie (H_i) eines Nutzers n_i gelöscht werden soll. Dieser hängt direkt von der angewandten Zielfunktion (engl. target function, *ta*) ab. Eine Zielfunktion bekommt eine Menge von Termen übergeben und liefert den Term mit dem geringsten Nutzen bezüglich *ta* zurück. Da es unterschiedliche Zielfunktionen gibt, gibt es folglich im Allgemeinen auch unterschiedliche anonymisierte Varianten S' eines Protokolls S. Diese haben den gleichen Grad der Privatheit, jedoch einen unterschiedlichen Nutzen. Im Folgenden bezeichnet U_{ta} den Nutzen eines anonymisierten Suchprotokolls. Der Nutzen eines anonymisierten Suchprotokolls ist die Summe der Einzelnutzen der Terme aller Historien bezüglich der gerade betrachteten Zielfunktion.

Beispiel 25: Sei S ein Suchprotokoll mit drei Nutzern und individuellen Historien für jeden Nutzer $H_1 = \{a,b,c\}$, $H_2 = \{a,b,d\}$ und $H_3 = \{b,c,d\}$. $a,\ldots,d$ sind Terme. ta zielt in diesem Beispiel auf Terme ab, auf die Werber hoch bieten. Sei $m = 2$, $k = 2$ der geforderte Grad der Privatheit. Es ist ersichtlich, dass das anonymisierte Protokoll bestehend aus $H'_1 = H'_2 = \{a,b\}$ und $H'_3 = \{b\}$ (2,2)-anonym ist, ebenso wie das bestehend aus $H'_1 = H'_3 = \{b,c\}$ und $H'_2 = \{b\}$. Bieten Werber jedoch für a hoch und für c niedrig, hätte die erste Variante einen höheren Nutzen.

Problembeschreibung: Gegeben unterschiedliche Grade der Privatheit und unterschiedliche Zielfunktionen ta und ta*, gibt es eine signifikante Differenz zwischen dem Nutzen eines anonymisierten Suchprotokolls, generiert unter Verwendung von ta, und dem Protokoll generiert mit ta*? Mit anderen Worten fragen wir, was ist die Auswirkung unterschiedlicher Zielfunktionen auf den Nutzen des Anonymisierungsergebnisses?

Wir können davon ausgehen, dass der größte Nutzen bezüglich ta^τ (U_{ta^τ}) entsteht, wenn wir auch die Zielfunktion ta^τ bei unserem Anonymisierungsalgorithmus einsetzen und nicht irgendeine andere. Es ist jedoch wichtig zu beachten, dass gierige (engl. greedy) Heuristiken nicht notwendigerweise auch zum besten Ergebnis führen.

Beispiel 26: Erweitert man Beispiel 25 um eine vierte Historie H_4={a,c,e}, so bilden H'_1={c}, H'_2={b,d}, H'_3={b,d} und H'_4={c} ein (2,2)-anonymes Protokoll (S'). Genauso, entsprechend bezeichnet als S'', erfüllt das H''_1={a,b,c}, H''_2={a,b}, H''_3={b,c} und H''_4={a,c}. Wir nehmen an, dass der Nutzen eines Terms $u = \text{bid} \cdot \text{clicks}$ ist und u_a=\$1, u_b=\$1.1, u_c=\$1.2, u_d=\$1.3, u_e=\$1.4. Für die erste Variante S' haben wir a anstelle von d entfernt und a anstelle von b, da a verglichen mit allen anderen Termen den geringsten Nutzen hat. Folglich hat die erste Variante einen Nutzen von $U_{\text{bid}*\text{click}}(S') = \$7.2$. Bei der zweiten Variante haben wir d entfernt, einen Term, der eigentlich einen hohen Nutzen hat. Das resultiert aber trotzdem in einem Gesamtnutzen von $U_{\text{bid}*\text{click}}(S'') = \$9.9$, das heißt einem Nutzen größer als dem bei S'.

6.3.3. (k,m)-Anonymität Algorithmus

In diesem Abschnitt beschreiben wir unseren Anonymisierungsalgorithmus (Algo-

Algorithmus 1 (k,m)-Anonymität - Greedy Heuristik

```
1:  S ⇐ Originales Suchprotokoll, S' ⇐ {};
2:  N := Gesamtzahl der Nutzer;
3:  fis := fpgrowth();
4:  while S != S' do
5:     S' ⇐ S
6:     for i = 1 to N do
7:        H_i ⇐ alle Terme von Nutzer n_i
8:        C_i ⇐ Kombinationen von Termen in H_i der Größe
9:        for j = 1 to |C_i| do
10:          c_ij := C_i[j]
11:          if support(fis, c_ij) < k then
12:             r := target_fct(c_ij, ta)
13:             H_i ⇐ H_i \ r
14:             fis ⇐ update_fis(H_i)
15:          end if
16:       end for
17:    end for
18:    S ⇐ ⋃ H_i
19: end while
```

rithmus 1). Dabei handelt es sich um eine Heuristik. S ist das Suchprotokoll, N ist die Anzahl Nutzer. H_i ist die Historie von Nutzer n_i und $S = \bigcup H_i$. fis umfasst die Menge aller häufigen Termkombinationen, also den Kombinationen, die mindestens von k Nutzern verwendet worden sind. Die Liste C_i beinhaltet alle Termkombinationen der Größe m, bestehend aus Termen in H_i. c_{ij} steht für die j-te Kombination in C_i. Wir geben mengenwertige Zuweisungen durch $\Leftarrow$ an.

Beispiel 27: Beispiel 25 mit $k = 2$ und $m = 2$ fortzusetzen würde bedeuten, fis enthält $\{a,b\}$, C_1 enthält $[\{a,b\},\{a,c\},\{b,c\}]$ und $c_{1,1}$ ist $\{a,b\}$.

Initial nutzen wir, um häufige Kombinationen von Termen zu finden (Zeile 3), die FP-Growth-Implementierung von [Bor05a]. Im nächsten Schritt bilden wir für alle Nutzer $n_i \in N$ alle möglichen Kombinationen von Termen aus H_i der Größe m und alle Teilmengen davon. Anschließend prüfen wir für jede dieser Kombinationen, ob sie Teilmenge von fis, der Menge der häufigen Termkombinationen, sind. Ist das für eine Termkombination der Fall, ist diese (k,m)-anonym. Ist der Support, das heißt die Häufigkeit des Terms, kleiner k ($support(c_{ij}) < k$), so müssen wir einen Term innerhalb der gerade betrachteten Kombination aus H_i löschen.

Unsere Heuristik ist gierig in dem Sinne, dass sie den Term löscht, der gemäß der Zielfunktion und der betrachteten Termkombination gerade am besten passt. Wir be-

Tabelle 6.3.: Auswahl des zu löschenden Terms

Iteration	Term	m	Häufigkeit	Status
1.	*eugene*	3	4	beibehalten
1.	*chicago*	3	2	Kandidat
1.	*smith*	3	2	Kandidat
2.	*chicago*	2	4	Kandidat
2.	*smith*	2	4	Kandidat
3.	*chicago*	1	6	beibehalten
3.	*smith*	1	4	löschen

rechnen den Term, der gelöscht werden soll, in Zeile 12 mit Hilfe der Funktion *target_fct*. *target_fct* kann sich auf drei verschiedene Arten verhalten. Diese sind 'random', 'fis' und mit Fokus auf die beschriebenen generischen und spezifischen Zielfunktionen. Die erste Alternative 'random' bedeutet, dass der Term, der aus c_{ij} gelöscht werden soll, zufällig ausgewählt wird. Wir nutzen dieses Verfahren als Richtwert für die Qualität unserer nach unterschiedlichen Zielfunktionen anonymisierten Daten. Die zweite Alternative 'fis' arbeitet direkt auf der Basis häufiger Elementmengen (frequent itemsets). Dieses Verhalten ist am nächsten verwandt mit existierenden Arbeiten zu mengenwertiger Anonymisierung. Wir löschen dabei den Term, der Teil der wenigsten Kombinationen in fis ist. Hat mehr als ein Term die minimale Häufigkeit, so berechnen wir die Häufigkeit des Terms in allen $m' = (m-1)$-elementigen Kombinationen. Wir wiederholen das, biss nur ein Term die minimale Häufigkeit hat oder $k = 0$ ist. Im letzten Fall wählen wir den zu löschenden Term zufällig aus der verbleibenden Menge aus.

Beispiel 28: Angenommen, es gibt, wie in Tabelle 6.3 beschrieben, die Kombinationen der Terme {eugene, chicago, smith} $\notin fis$. Folglich müssen wir einen der drei Terme löschen. Wir berechnen also, wie oft jeder Term in häufigen Elementmengen existiert. In unserem Beispiel kommt 'eugene' viermal, 'chicago' und 'smith' jeweils zweimal vor (Spalte 'Häufigkeit'). Entsprechend behalten wir 'eugene' und löschen einen der verbleibenden Terme. Wir verringern m. Erneut kommen jedoch 'chicago' und 'smith' gleich oft (viermal) vor. Wir verringern m erneut. 'smith' ist seltener, das heißt wird gelöscht.

Abschließend beschreibt Algorithmus 2 die dritte Alternative. Hier ist der Nutzen eines Terms entsprechend der Zielfunktion ta definiert. Der Algorithmus gibt den Term zurück, dessen Nutzen bezüglich ta minimal ist (Zeile 6). Haben zwei Terme den gleichen Nutzen, wählen wir einen zufällig aus.

Unsere Implementierung unterstützt $ta \in$ {random, # Terme, # Nutzer, max Gebot, # Klicks, # Werbeeinblendungen, Umsatz}. '# Terme' beschreibt die Größe des Suchprotokolls, '# Nutzer' die Anzahl der Nutzer, für die die Suchhistorie anonymisiert

Algorithmus 2 target_fct(c,ta)

```
1: c := Termkombinationen der Größe m
2: v ⇐ {}; /*Term Nutzen*/
3: for i = 1 to m do
4:     v ⇐ get_term_utility(i, ta);
5: end for
6: return gib zufälliges Element mit Nutzen = v.min() zurück;
```

werden kann, ‘max Gebot’ die Gebote, ‘# Klicks’ die Anzahl Klicks auf eine Werbung, ‘# Werbeeinblendungen’ die Anzahl Werbeeinblendungen und ‘Umsatz’ wie beschrieben einen Hinweis auf den möglichen Umsatz.

Nach dem Löschen eines Terms müssen wir fis aktualisieren (Algorithmus 1, Zeile 14). Wir verringern den Support aller Termkombinationen, die diesen Term beinhalten. Wichtig ist dabei, nur die Termkombinationen zu betrachten, die man aus Termen $t \in H_i$ bilden kann. Wir löschen die Kombinationen, die anschließend einen neuen Support $\leq k$ haben.

Unser Algorithmus bearbeitet einen Nutzer nach dem anderen. Folglich kann das Löschen von Termen und (ehemals) häufiger Termkombinationen die Historien schon bereits verarbeiteter Nutzer betreffen. Aus diesem Grund wenden wir unsere Heuristik solange iterativ an, bis kein Term mehr gelöscht werden muss.

6.4. Evaluation

In diesem Abschnitt beantworten wir unsere Forschungsfrage. Wir beschreiben zuerst, welche Daten wir in der Evaluierung nutzen (Abschnitt 6.4.1). Danach, auch wenn die Effizienz unserer Heuristik nicht unser eigentliches Ziel ist, geben wir eine Intuition des Laufzeitverhaltens unseres Algorithmus an (Abschnitt 6.4.2). Wir beschreiben detailliert, welche Methoden wir zur Bewertung des Nutzen des anonymisierten Suchprotokolls für CSE-Nutzer und Anbieter einsetzen (Abschnitt 6.4.3). Ergebnisse unseres Vergleichs unterschiedlicher Zielfunktionen und Werte für k beschreiben wir in Abschnitt 6.4.4. Abschnitt 6.4.5 zielt auf den Einfluss des Parameters m ab, das heißt die Größe der Termkombination, ab der eine Termkombination möglicherweise zu der Identifikation einer Person führt.

6.4.1. Suchprotokolle und Marketingdaten

In diesem Abschnitt stellen wir kurz die in der Evaluation verwendeten Daten vor. Das sind einerseits das Suchprotokoll von Altavista und andererseits die Marketingdaten von Yahoo!.

Altavista Suchprotokoll. Das Altavista-Suchprotokoll wurde 2002 publiziert[6]. Es umfasst 3,5 Millionen Anfragen von einem Tag, abgesetzt von 370.585 Nutzern. Das Protokoll hat das Schema `Q=<user_id, query, timestamp>`

Wir haben die folgenden Vorverarbeitungsschritte angewandt: (i) Unser Ziel ist es, identifizierende Information zu anonymisieren, nicht aber Anfragen von Metasuchmaschinen[7], Crawlern[8] etc. Kann ein potentieller Angreifer einen solchen Dienst in einem anonymisierten Suchprotokoll identifizieren, so würde das mit jedem identifizierten Dienst die Ununterscheidbarkeit der Nutzer von k auf $k-1$ verringern. Folglich haben wir Anfragen von nicht-menschlichen Nutzern gelöscht. Dazu haben wir die (sehr einfache) Methode von [JSP05] angewandt, bei der Nutzer mit Such-Sitzungen (engl. session), die mehr als 100 Anfragen enthalten, entfernt werden. (ii) Außerdem normalisieren wir unseren Datensatz. Beispielsweise unterscheiden sich die beiden Anfragen 'ad–placement' und 'ad placement' nur in einem Sonderzeichen. Unsere Vorverarbeitung entfernt alle Zeichen außer [A-Za-z0-9À-ÖÙ-Ýß-öù-ž]. (iii) Privatheitsbedrohungen können aus der Kombination von Termen aus unterschiedlichen Anfragen der gleichen Person entstehen. Wir extrahieren die Terme aus jeder Anfrage und speichern sie in einer Relation mit dem Schema `QT=<user_id, query_id, term_id>`

Unsere Vorverarbeitung führt zu 1.846.134 Anfragen von 367.803 Nutzern. Die Anfragen umfassen 251.115 voneinander unterschiedlichen Terme und 5.501.825 Vorkommen dieser Terme. Merke, dass die Anzahl Terme weit über der Anzahl Terme konventioneller Wörterbücher liegt, zum Beispiel für die deutsche oder englische Sprache. Das liegt an der Verwendung von Eigennamen ('myspace'), Fehlerkennzahlen ('oraXXXX'), etc.

Yahoo! Marketingdaten. Wir sind aus Sicht der Anbieter an dem Wert anonymisierter Suchprotokolle zu Werbezwecken interessiert. Zu diesem Zweck ordnen wir jedem Term einen Wert zu. Die Yahoo! Marketingseite[9] bietet, gegeben einen Term, eine Abschätzung an für (i) die Anzahl von `Werbeeinblendungen`, (ii) die Häufigkeit, mit der eine Anzeige angeklickt wird (`# Klicks`) und (iii) das maximale Gebot (`max Gebot`), das Yahoo! seinen Werbern empfiehlt (unter Berücksichtigung eines täglichen Werbebudgets). Wir haben die Seite systematisch mit Hilfe eines Programms ausgelesen und die Marketinginformationen für jeden der ungefähr 250.000 unterschiedlichen Terme aus dem Altavista-Suchprotokoll gesammelt.

Korrelationsanalyse. Um unsere Frage bezüglich des Einflusses der Zielfunktion auf

[6] Altavista-Suchprotokoll 2002, bereitgestellt von Jim Jansen (jjansen@acm.org)

[7] Eine Metasuchmaschine hat keinen eigenen Index, sondern setzt eine Suchanfrage an mehrere Suchmaschinen gleichzeitig ab und kombiniert die Ergebnisse der Antworten zu einem Suchergebnis.

[8] Crawler sind kleine Programme, die das Internet durchsuchen.

[9] `http://sem.smallbusiness.yahoo.com/searchenginemarketing/marketingtool.php` Juni 2010

Tabelle 6.4.: Pearson-Korrelationsanalyse der Ausgangsdaten

		Umsatz	# Einbl.	# Klicks	max Gebot	# Nutzer.
# Terme	corr.	.002	.111*	.024*	.001	.761*
	sig.	.326	.000	.000	.348	.000
	N	64,694	117,916	64,694	117,916	117,916
# Nutzer	corr.	.003	.151*	.031*	.002	
	sig.	.259	.000	.000	.254	
	N	64,694	117,916	64,694	117,916	
max Gebot	corr.	.579*	.002	-.003		
	sig.	.000	.296	.225		
	N	64,694	117,916	64,694		
# Klicks	corr.	.040*	.400*			
	sig.	.000	.000			
	N	64,694	64,694			
# Einbl.	corr.	.020*				
	sig.	.000				
	N	64,694				

das Ergebnis der Anonymisierung beantworten zu können, ist es wichtig, die Charakteristiken der Eingabedaten genau zu kennen. Zu diesem Zweck berechnen wir die paarweise Korrelation zwischen allen relevanten Variablen, die wir in unserer Evaluation untersuchen. Diese Variablen sind die Häufigkeit eines Terms im Suchprotokoll, die Anzahl Nutzer, das maximale von Yahoo! empfohlene Gebot, die Abschätzung für die Klicks, die Anzahl Werbeeinblendungen und der mögliche Umsatz. Für dichotome Variablen, hier die Variable {im Log, nicht im Log} (unterschiedliche Terme) ist eine Korrelationsanalyse nicht möglich [Bor05b].

Tabelle 6.4 zeigt die Pearson-Korrelation (corr.), die Signifikanz der Korrelation (sig.) und die die Anzahl Terme (N), die wir bei der Korrelationsanalyse berücksichtigt haben. Da Yahoo! für 45% der Terme in unserem Altavista-Suchprotokoll keine Abschätzungen für die Anzahl Klicks anbietet, haben alle paarweisen Kombinationen mit ‘Klicks’ ein kleineres N. Ergebnisse, die mit einem ‘*’ gekenntzeichnet sind, sind signifikant bei einem Signifikanzniveau von 0,01 (einseitiger Test). Für 9 Kombinationen sehen wir eine signifikante Korrelation, umgekehrt ist das bei 6 Kombinationen nicht der Fall. Alle Kombinationen bis auf eine (# Klicks, max Gebot) haben eine positive Korrelation. Daraus schließen wir, dass in vielen Fällen die Fokussierung einer Zielfunktion auf eine der Variablen auch einen positiven Effekt auf die Anonymisierungsergebnisse unter Verwendung der anderen Zielfunktionen hat. Das hängt jedoch stark von der Variablen ab. Für die Variable ‘max Gebot’ besteht die einzige signifikante Korrelation mit der Variable ‘Umsatz’. Das ist nicht überraschen, da ‘max Gebot’ Einfluss auf die Variable ‘Umsatz’ hat. Die Variable ‘# Klicks’ hat hingegen eine signifikante Korrelation mit vier anderen Variablen. Wir werden die Auswirkungen dieser

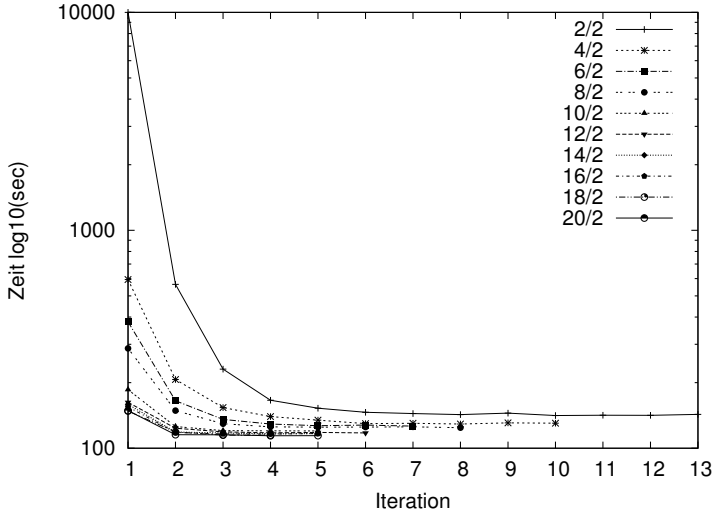


Abbildung 6.1.: m=2: Zeit / Iteration

Korrelationen in Abschnitt 6.4.3 analysieren.

6.4.2. Laufzeitverhalten

Unser Ziel ist der Vergleich unterschiedlicher Zielfunktionen, nicht aber das Erreichen einer optimalen Berechnungszeit. Wir werden im Folgenden jedoch zumindest eine Intuition des Laufzeitverhaltens unseres Algorithmus für unterschiedliche k und m geben.

Wir setzen Standardhardware (Dual-Core AMD Opteron 2218), 8GB RAM, Java und eine einfädrige Implementierung ein. Außerdem nutzen wir eine relationale Datenbank, um die Suchhistorien während des Anonymisierungsprozesses zu speichern. Beispielsweise protokollieren wir für jeden Grad von Privatheit und für jede Iteration alle Löschungen von Termen, die unser Algorithmus vornimmt. Die Interaktion mit der Datenbank hat den stärksten Einfluss auf die Laufzeit.

Vergleicht man die Laufzeiten der Berechnungen für die unterschiedlichen Zielfunktionen, so sieht man, dass die Zielfunktion, die die meiste Zeit benötigt, die ist, die nur auf den häufigen Elementmengen arbeitet (fis, der zweite Fall in Abschnitt 6.3.3). Die Zeit für den schwierigsten Fall, das heißt kleine k, ist besonders interessant. Je größer k, desto kleiner ist das Vokabular (Anzahl unterschiedlicher Terme) und folglich die für die Berechnung erforderliche Zeit.

Die erste Iteration für jedes k erfordert das Löschen einer großen Menge seltener Terme und Termkombinationen (siehe auch Anhang E.1). Entsprechend erfordert die erste Iteration die meiste Zeit, zum Beispiel 02h:55m:58s für $k = 2$. Mit der zweiten Iteration sinkt die erforderliche Zeit bereits auf 09m:25s, das heißt ungefähr um Faktor

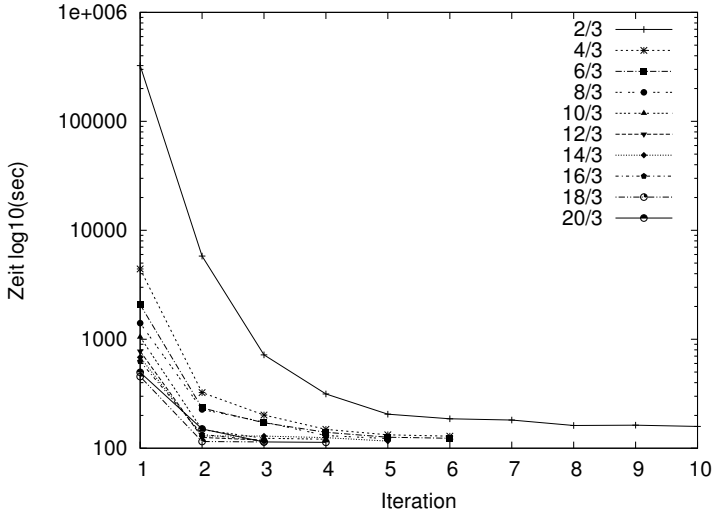


Abbildung 6.2.: m=3: Zeit / Iteration

17. Für $k > 2$ ist die Zeit auch für die erste Iteration deutlich reduziert, zum Beispiel auf 7m:41s für $k = 5$. Für $m = 2$ beträgt die mittlere Zeit pro Iteration 4 Minuten und die mittlere Anzahl der Iterationen 6,7.

Qualitativ ausgedrückt hält diese Beobachtung auch für (k,3)-Anonymität. Trotzdem ist die erste Iteration mit 90h:08m:24s für $k = 2$ und ausgehend vom initialen Suchprotokoll deutlich kostenintensiver als für $m = 2$ (Faktor 31). Das liegt an der Anzahl möglicher Kombinationen. Das größte T_i enthält 493 Terme. Entsprechend Algorithmus 1 und $m = 3$ müssen wir dafür 20.092.215 Kombinationen auf ausreichenden Support prüfen. Startet man hingegen von (2,2)-Anonymität, so können wir (2,3)-Anonymität innerhalb von 06h:35m:22s, das heißt in 7% der ursprünglich erforderlichen Zeit, berechnen. Auch hier reduziert sich die Zeit für die erste Iteration, wenn $k > 2$ gilt. Für $k = 5$ benötigt die erste Iteration 48m:45s. Die mittlere Zeit pro Iteration ist 13 Minuten, mit im Mittel 4,8 Iterationen.

6.4.3. Nutzen des anonymisierten Suchprotokolls

In diesem Abschnitt analysieren wir zuerst die generischen Eigenschaften des anonymisierten Suchprotokolls, das heiß, inwieweit CSE-Nutzer trotz Anonymisierung noch von CSEs profitieren können. Wir untersuchen dazu, wie viele unterschiedliche Terme nach der Anonymisierung im Suchprotokoll verbleiben, wie groß das Suchprotokoll noch ist und für wie viele Nutzer eine Anonymisierung der Suchhistorie möglich ist.

Anschließend fokussieren wir auf die für die Anbieter relevanten Eigenschaften. Das umfasst das maximale Gebot, das Yahoo! für einen Term empfiehlt, die Anzahl Klicks

Tabelle 6.5.: Nutzen der anonymisierten Daten

m	k	untersch. Terme	# Terme	# Nutzer	max Gebot	# Klicks	# Einblendungen	Umsatz
2	2	10.88	45	95	51	85	85	61
	10	2.12	29	87	33	75	71	41
	20	1.09	26	83	29	71	67	36
	40	0.56	23	78	25	68	63	33
	100	0.18	19	70	20	61	58	27
3	2	10.86	37	95	42	78	76	55
	10	2.12	25	87	31	71	67	43
	20	1.11	23	83	28	68	64	39
	40	0.56	21	78	24	65	62	35
	100	0.18	18	70	19	61	57	30

auf eine Werbeeinblendung, die Anzahl der Einblendungen und den möglichen Umsatz.

Bis auf die Anzahl unterschiedlicher Terme haben wir für jede dieser Eigenschaften Zielfunktionen entwickelt. Für jede Zielfunktion in Kombination mit $m = 2$ als auch $m = 3$ und mit k zwischen 2 und 100, berechnen wir das anonymisierte Suchprotokoll. In allen Fällen hat die jeweilige Zielfunktion auch zu dem besten Ergebnis bezüglich ihres Fokus geführt. So hat, verglichen mit dem Einsatz aller anderen Zielfunktionen, das Abzielen auf ein großes Suchprotokoll auch tatsächlich zu dem größten anonymisierten Protokoll geführt. Für die unterschiedlichen Terme im Protokoll zeigen wir jeweils die besten Messergebnisse der anderen Zielfunktionen an.

Tabelle 6.5 zeigt unsere Ergebnisse. Die obere Hälfte sind die Ergebnisse für $m = 2$ und $k = 2, 10, 20, 40, 100$, die untere Hälfte die für $m = 3$.

Unterschiedliche Terme Die Spalte 'untersch. Terme' zeigt, dass selbst für die niedrigsten Werte $m = 2, k = 2$, das heißt die schwächste Anonymisierung, die Anzahl unterschiedlicher Terme signifikant verringert wird, hier auf 10,88%. Für den höchsten gemessenen Grad der Privatheit ($m = 3, k = 100$) sind das sogar nur 0.18%. Der Hauptgrund dafür ist der Long-Tail-Effect von Suchprotokollen. Das bedeutet, dass die Nutzer sehr viele Terme selten und sehr wenige häufig verwenden. Trotzdem verbleiben nach der stärksten gemessenen Anonymisierung 500 unterschiedliche Terme, die CSE-Nutzer anonym austauschen und die Werber zumindest für Werbung zu einer relativ breiten Menge von Themen nutzen können.

 # Terme Mit $m = 2, k = 2$ messen wir für die Größe des anonymisierten Protokolls eine signifikante Verkleinerung um 55%. Auch dieses Verhalten lässt sich wieder auf den Long-Tail-Effekt zurückführen. Für $m = 2, k = 20$ können wir noch 26% der Protokollgröße wahren, für $m = 3, k = 100$ 18%. Das Ergebnis bedeutet, dass selbst wenn wir nur zwischen 11% und 0,18% unterschiedlicher Terme anonymisieren können, die verbleibenden Terme häufig sind. Somit besteht ein Interesse von CSE-Nutzern, diese Terme auszutauschen, da häufige Terme wahrscheinlich auch relevant sind. Ebenso besteht ein Interesse von Werbern, auf solche Terme zu bieten.

 # Nutzer Unabhängig von m und für kleine k beinhaltet das Anonymisierungsergebnis 95% der Nutzer. 5% mussten zum Erreichen unserer Definition von (k,m)-Anonymität entfernt werden. Und auch für große k, zum Beispiel $k = 100$, können wir von 70% aller Nutzer mindestens einen Teil ihrer Suchhistorie anonymisieren. Wir halten diese Zahlen für vielversprechend, um zumindest einige Terme anderen CSE-Nutzern zur Verfügung stellen zu können und um für Anbieter grob die Interessen der Nutzer abzuleiten.

Nach den generischen Maßen wenden wir nun die Maße an, die speziell für Werbung geeignet sind.

max Gebot Unsere Ergebnisse (Spalte 'max Gebot') zeigen, dass nach der Anonymisierung für $m = 2$, $k = 2$, die Summe der Gebote auf Terme, die im Anonymisierungsergebnis beinhaltet sind, immer noch 51% der Ausgangssumme umfasst. Für $m = 3$ und $k = 100$ ist dies immerhin noch 20%.

 # Klicks Für große $k = 100$ und $m = 3$ führen die Terme im Suchprotokoll zu 61% der abgeschätzten Klicks, verglichen mit dem Ausgangsdatenbestand. Für $m = 2$, $k = 2$ liegt diese Rate sogar bei 85%.

 # Werbeeinblendungen Unsere Ergebnisse (Spalte '# Einblendungen') sind sehr ähnlich denen für Klicks. Wir können die Terme im Suchprotokoll halten, die zu 57% bis 85% aller Werbeeinblendungen führen, abhängig vom Grad der Privatheit.

Umsatz Der Umsatz kombiniert das Gebot der Werber, die voraussichtliche Anzahl Klicks der Nutzer und die Termhäufigkeit. Unsere Ergebnisse zeigen (Spalte 'Umsatz'), wieder abhängig vom Grad der Privatheit, dass die Umsätze zwischen 30% und 61% variieren.

6.4.4. Einfluss der Zielfunktion

In diesem Abschnitt vergleichen wir die unterschiedlichen Zielfunktionen. Tabelle 6.6 enthält unsere Ergebnisse. Für jedes Maß geben wir die Differenz der besten erreichten Ergebnisse verglichen mit (i) 'random' (Spalte 'R'), (ii) dem schlechtesten Ergebnis (Spalte 'W') und (iii) der Funktion, die nur auf häufigen Termkombinationen arbeitet (fis), an. 'Random' ist der Mittelwert von zwei Läufen.

Unser erstes Ergebnis ist, dass für die generischen Charakteristiken die Differenz

Tabelle 6.6.: Einfluss der Zielfunktionen (%)

m	k	untersch. Terme			# Nutzer			# Terme			max Gebot			# Klicks			# Einblend.			Umsatz		
		R	W	fis	R	W	fis	R	W	fis	R	W	fis	R	W	fis	R	W	fis	R	W	fis
2	2	0.00	4.16	4.08	13	2	0	2	0	0	20	13	12	28	15	8	31	9	6	20	5	5
	10	0.00	0.92	0.90	11	2	0	12	1	0	17	13	11	34	28	13	32	14	8	16	5	5
	20	0.00	0.40	0.39	11	2	1	17	1	0	16	12	11	35	33	15	32	15	9	16	5	5
	40	0.00	0.16	0.15	11	2	1	21	2	0	14	11	9	38	36	18	31	16	9	15	5	5
	100	0.00	0.00	0.14	11	2	1	28	2	0	13	9	7	37	37	19	30	16	10	15	5	5
3	2	0.00	4.18	4.07	10	3	0	2	0	0	15	14	11	32	31	12	29	16	7	21	12	10
	10	0.00	0.88	0.77	10	2	0	12	1	0	16	14	11	39	40	21	34	19	11	22	12	12
	20	0.00	0.31	0.27	10	2	1	17	1	0	15	13	11	40	41	24	34	19	12	21	12	12
	40	0.00	0.16	0.14	10	2	1	21	2	0	14	11	9	40	41	25	33	20	12	19	11	11
	100	0.00	0.02	0.00	10	2	1	28	2	0	13	9	7	39	40	25	32	19	12	19	10	10

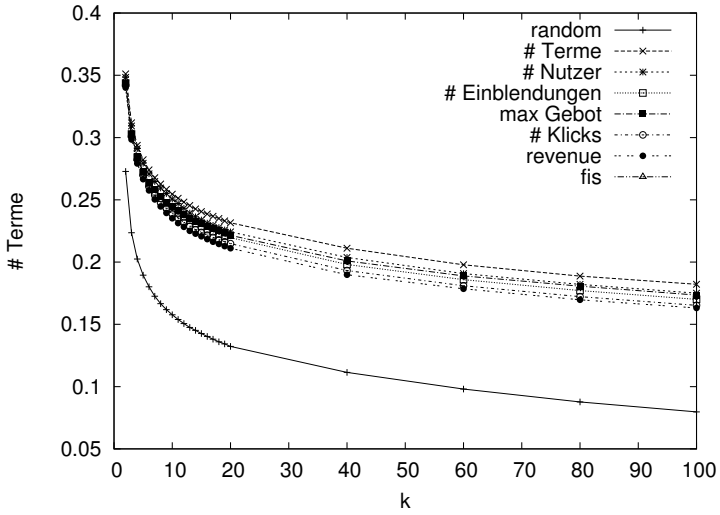


Abbildung 6.3.: Protokollgröße / Zielfunktion

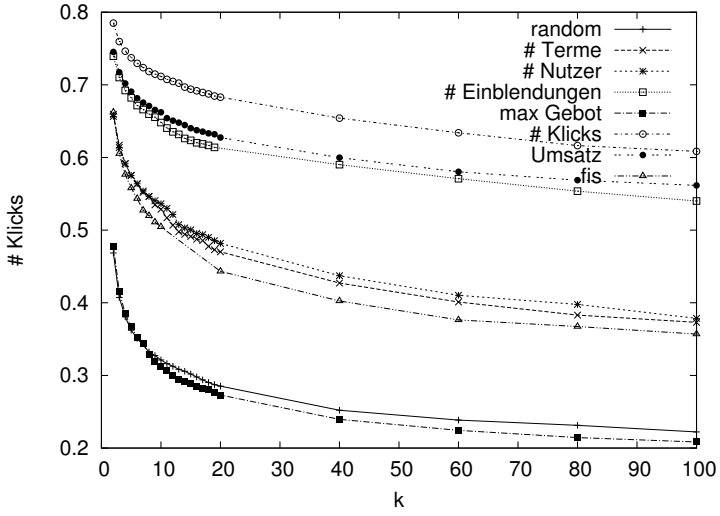


Abbildung 6.4.: Klicks / Zielfunktion

zwischen der besten Zielfunktion und der schlechtesten in allen Kombinationen kleiner 4,2% ist.

Abbildung 6.3 zeigt exemplarisch die resultierende Größe des Suchprotokolls für die unterschiedlichen Zielfunktionen und $m = 3$, $k = 2 \ldots 100$ (weitere Abbildungen für $m = 2,3$ und unterschiedliche Zielfunktionen befinden sich in Anhang E.2). Im Vergleich zu 'random' ist dort die Differenz von bis zu 28% relativ hoch. Folglich können wir erkennen, dass trotz der kleinen Differenz zwischen den einzelnen generischen Zielfunktionen ein klarer Unterschied zu der Verwendung keiner wohlüberlegten Zielfunktion besteht. Zielt man mit der Zielfunktion auf viele Nutzer ab, so führt das zu dem zweitgrößten Protokoll. Das liegt an der hohen Korrelation zwischen der Anzahl Nutzer, die nach einem Term gesucht haben, und der Häufigkeit eines Terms (0,761 gemäß Tabelle 6.4).

Unser zweites Ergebnis betrifft die für Werbung spezifischen Charakteristiken. Hier ist der Gesamtnutzen bezüglich einer Charakteristik stark abhängig von der Zielfunktion, das heißt, die Differenz zwischen dem besten und dem schlechtesten Ergebnis ist signifikant. Am Beispiel der abgeschätzten Klicks auf eine Anzeige (Abbildung 6.4) sieht man, dass die Wahl der besten Zielfunktion (# Klicks) und der schlechtesten Zielfunktion (max Gebot) einen Unterschied von mehr als 40% ausmacht. Das liegt darin begründet, dass die Variablen 'Klicks' und 'max Gebot' (wenn auch nicht signifikant) negativ korreliert sind (Tabelle 6.4). Interessanterweise ist die beste Zielfunktion für 'unterschiedliche Terme' (wie gesagt, haben wir dafür keine eigene Zielfunktion entwickelt) das zufällige Löschen von Termen. Ein weiterer Indikator dafür, dass die Zielfunktion signifikanten Einfluss auf den Nutzen eines Suchprotokolls nach der Anonymisierung hat.

6.4.5. Einfluss des Grades der Privatheit

Eine wichtige Entscheidung der Instanz, die die Daten anonymisiert, ist die Wahl von m. m ist die Größe der Termkombinationen, die potentiell zu einer Re-Identifikation führen können. Wir erinnern uns, dass $m = 2$ im AOL-Fall vor der Re-Identifikation geschützt hätte (Abschnitt 6.2.4). Wir evaluieren unsere Ergebnisse für $m = 2$ aber auch für $m = 3$. Überraschenderweise ist der Effekt von m im Vergleich zu dem von k minimal. Für $k \geq 10$ ist die Differenz des Ergebnisses zwischen $m = 2$ und $m = 3$, gleich welche Zielfunktion genutzt wurde, kleiner 4%. Für $k = 100$ sind die meisten Werte sogar gleich. Für die generischen Zielfunktionen sind die Unterschiede unter einem Prozent. Die größte Differenz von 7% liegt bei 'Umsatz' vor. Wir schließen daraus, dass für große k ein Wert für $m < 3$ zu einem unnötig niedrigen Grad der Privatheit führen würde, ohne einen positiven Effekt auf den Nutzen.

6.5. Zusammenfassung

Die Anonymisierung von Suchprotokollen ist aus mehrfacher Hinsicht wünschenswert: CSE-Nutzer können, wie von ihnen gefordert, anonym kollaborieren. Außerdem schützt Anonymisierung die Privatheit der Nutzer gegenüber dem Suchmaschinenanbieter. Für den Anbieter ist die Haupteinnahmequelle personalisierte Werbung. Der Umsatz hängt dabei von der Möglichkeit ab, die Nutzerintention vorherzusagen. Eine Verbesserung dieser Vorhersage wird voraussichtlich durch das Erheben, Speichern und Verarbeiten umfangreicher Suchprotokolle erfolgen. Das wiederum setzt die Privatheit der Nutzer aufs Spiel, generiert aber auch für die Anbieter eine Vielzahl von Möglichkeiten, gegen geltendes Datenschutzrecht zu verstoßen. Die Anbieter können durch Anonymisierung die Privatheit ihrer Nutzer schützen und zugleich Daten ohne die Erfüllung oftmals kostenintensiver Auflagen des Gesetzgebers verarbeiten. Die Voraussetzung für beide, Nutzer wie Anbieter, ist, dass auch ein anonymisiertes Suchprotokoll seinen ursprünglichen Nutzen erfüllt: Nutzer müssen sich nach wie vor gegenseitig beim Suchen unterstützen und Anbieter Werbung schalten können.

In diesem Kapitel haben wir gezeigt, dass CSE-Nutzer eine Vielzahl von Termen auch anonymisiert austauschen können und dass Anbieter, trotz eines Verzichts auf identifizierende Informationen, effektiv Werbung schalten können. Wir verwenden für unseren Ansatz die Definition von (k,m)-Anonymität für mengenwertige Daten. Wir haben einen Algorithmus implementiert, der eine flexibel Wahl der Zielfunktion unterstützt. Ein Ziel für CSE-Nutzer ist beispielsweise der Erhalt großer anonymisierter Suchprotokolle, ein Ziel für die Anbieter sind Suchprotokolle mit solchen Termen, die zu vielen Klicks auf Werbung führen.

Mittels einer umfangreichen Evaluierung auf Realweltdaten haben wir gezeigt, dass anonymisierte Suchprotokolle wertvolle Informationen für CSE-Nutzer als auch Suchmaschinenanbieter und Werber beinhalten. Ein Ausschnitt unserer Ergebnisse ist, dass die von uns genutzte Anonymisierung, abhängig vom Grad der Privatheit, Daten zu 70% bis 95% aller Nutzer wahren kann. Bezüglich des heute wichtigsten Geschäftsmodells von Anbietern im Netz – pay-per-click – haben wir gezeigt, dass wir alternativ auch solche Terme bei der Anonymisierung wahren können, die zu 61% bis 85% der Klicks auf Werbeanzeigen führen. Diese Berechnung basiert auf Marketingdaten von Yahoo!. Wie sich herausgestellt hat, hat die Wahl von $m = 2$ oder $m = 3$ einen nahezu vernachlässigbaren Effekt auf das Anonymisierungsergebnis – unabhängig von der verwendeten Zielfunktion. Die Zielfunktionen selbst sind ausgesprochen wichtig: Wählt man zum Beispiel eine unglückliche Zielfunktion bei dem eigentlichen Ziel, die Klicks auf Werbung hochzuhalten, so kann das zu 40% weniger Klicks führen als bei der Wahl einer geschickten Zielfunktion. Das Bedeutet auch, dass eine Anonymisierung, die gleichzeitig einen Nutzen für CSE-Nutzer und Anbieter haben soll, einer eigenen Zielfunktion bedarf, die die jeweiligen Ziele sinnvoll kombiniert. Eine Ausnahme ist, wenn beide Parteien auf generische Charakteristiken abzielen. Ist das Ziel

beispielsweise zum Einen die Größe des anonymisierten Protokolls und zum Anderen die Anzahl der Nutzer im Log, so hat sich gezeigt, dass die Differenz der Nutzen der anonymisierten Daten trotz des Einsatzes unterschiedlicher Zielfunktionen kleiner 4,2% ist.

Während diese Ergebnisse aus unserer Sicht an sich schon vielversprechend sind, so kann man davon ausgehen, dass größere Suchprotokolle zu noch besseren Ergebnissen führen. Dieser Aussage liegt die Annahme zugrunde, dass mit einer steigenden Zahl Nutzer die Anzahl der seltenen Terme nicht gleichermaßen (stark) mitwächst und somit mehr Terme häufig werden. Außerdem können wir annehmen, dass das von den Nutzern geforderte k, also die Anzahl der Nutzer, von denen sie ununterscheidbar sein möchten, ebenfalls nicht mit der Anzahl Nutzer wächst. Ununterscheidbar von 100 oder gar 1000 Nutzern zu sein erscheint sinnvoll, unabhängig davon, ob das Suchprotokoll eine Million Nutzer oder zehn Millionen Nutzer umfasst.

Unsere Schlussfolgerung ist, dass zum Schutz der Nutzerprivatheit und der Entbindung von Suchmaschinenanbietern von rechtlichen Anforderungen bei einer gleichzeitig gewahrten Werbemöglichkeit, Anonymisierung einen vielversprechenden Ansatz darstellt.

7. Beitrag der Arbeit und Ausblick

Das Internet hat den Alltag durchdrungen. Anbieter stellen eine Vielzahl von Diensten für alle Schichten der Gesellschaft bereit. Mit dem Web 2.0 sorgen sogar die Nutzer kollaborativ für die inhaltliche Ausgestaltung der Dienste. Zur Registrierung bei einem Dienst oder bei dessen Nutzung geben Personen umfangreich sensible, personenbezogene Daten preis – eine Bedrohung für die Privatheit.

Es existieren Technologien zum Schutz der Privatheit, sogenannte Privacy-Enhancing Technologies (PET). Die große Anzahl berichteter Bedrohungen für die Privatheit und von Datenschutzverstößen zeigt, dass die existierenden PETs nicht ausreichen oder nicht so eingesetzt werden können, dass Nutzer damit ihre Privatheitspräferenzen durchsetzen können.

In dieser Arbeit haben wir anhand von Nutzerstudien untersucht, welche Privatheitspräferenzen Nutzer spezifizieren. Außerdem haben wir mehrere PETs implementiert und Personen im Alltag verwenden lassen. Wir haben dadurch Erkenntnisse gewonnen, wie ein PET geartet sein muss, damit Nutzer ihre Privatheitspräferenzen ausdrücken können. Weiter haben wir aus unseren Studien Anforderungen an PETs abgeleitet, damit Nutzer mit einem PET die Privatheitspräferenzen auch durchsetzen können.

Außerdem haben wir erkannt, dass Nutzer ein Vertrauensproblem mit der Datenschutzpraxis der Anbieter haben. Wir haben aus diesem Grund in einer umfangreichen Analyse der Anbieter untersucht, ob diese eine Vielzahl von Datenschutzverstößen begehen, die das niedrige Vertrauen rechtfertigen. Darauf aufbauend haben wir selbst einen Ansatz zur Identifikation von Datenschutzverstößen entwickelt. Der Ansatz basiert auf der Kollaboration von Nutzern und auf in Software gegossener juristischer Expertise zur Identifikation von Datenschutzverstößen.

Zuletzt haben wir einen Weg basierend auf Anonymisierung aufgezeigt, der Nutzern wie Anbietern hilft: Er schützt die Privatheit der Nutzer, er entbindet die Anbieter von komplexen und oftmals teuren Anforderungen der Gesetzgebung, und er versucht gleichzeitig, den Nutzen zu wahren, den die personenbezogenen Daten für den Anbieter haben.

7.1. Beitrag

Die Ergebnisse dieser Arbeit sind aus dreierlei Hinsicht interessant.

Erstens haben wir für unterschiedliche, populäre Web 2.0-Anwendungen Privat-

heitspräferenzen von Nutzern identifiziert. Diese haben wir auf eine einfache Struktur abbilden können. Weiter haben wir aus diesen Ergebnissen Anforderungen an PETs abgeleitet, die es Entwicklern zukünftiger PETs erlauben, diese zielgerichtet auf die Privatheits- und Nutzerpräferenzen hin zu entwickeln. Die Anforderungen beziehen sich auf (i) den Inhalt einer Präferenz, das heißt 'wer', 'wann', 'was' sehen darf, auf (ii) Anforderungen an die Benutzbarkeit und auf (iii) funktionale Anforderungen. Letzteres beschreibt insbesondere 'wie' die Anforderungen aus (i) und (ii) realisiert werden sollten. Ein zentrales Ergebnis ist, dass PETs, zum Beispiel ein 'Privat'-Schalter, die ein kontinuierliches Bewusstsein vom Nutzer fordern, versagen. Das sind aber genau die PETs, die aktuell bei standortbezogenen Diensten wie Google Latitude installiert sind. Außerdem sind Nutzer durchaus bereit, komplexen, schwer verständlichen aber automatisch arbeitenden PETs zu vertrauen, wenn sie in Kombination mit besagtem Schalter angeboten werden, die Nutzer also die letzte Entscheidung darüber haben, ob Daten preisgegeben werden.

Zweitens haben wir für eine breite Menge von Online-Diensten eine Vielzahl von Datenschutzverstößen nachgewiesen. In diesem Zusammenhang haben wir insbesondere die Hypothese eines Vollzugsdefizits im Datenschutz belegt. Wir haben erkannt, dass weder der einfache Internetnutzer, die Unternehmen, noch die Datenschutzaufsichtsbehörden die Verstöße effizient und systematisch identifizieren können. Dies liegt auf der einen Seite an mangelnder Kompetenz und auf der anderen Seite an fehlenden Ressourcen. Mit Hilfe eines von uns entwickelten Ansatzes, bei dem wir die erforderliche Expertise zur Identifikation von Datenschutzverstößen mit Software abgebildet haben, haben wir dem entgegengewirkt. Evaluiert mittels einer Nutzerstudie können Personen ohne Datenschutzexpertise über 80 Prozent der Verstöße identifizieren, die auch Experten finden.

Drittens hat sich unser auf Anonymisierung basierender Ansatz als vielversprechend erwiesen. Im Kontext von Suchprotokollen haben wir gezeigt, dass anonymisierte Suchprotokolle immer noch solche Daten beinhalten, die Nutzer bei kollaborativen Suchmaschinen austauschen und Anbieter von Suchmaschinen zum effektiven Schalten von Werbung einsetzten können – trotz Schutz der Privatheit und der Entbindung des Anbieters von datenschutzrechtlichen Anforderungen.

7.2. Ausblick

Im Folgenden geben wir einen kurzen Ausblick über mögliche interessante Folgearbeiten. Wir verweisen jeweils auf den Kontext dieser Arbeit, zu dem wir einen Vorschlag machen.

Reziprozität Teilnehmer der Studie zu kollaborativen Suchmaschinen haben Strategien formuliert, die sich auf Strategien anderer Nutzer beziehen, das heißt reziproke Strategien. Es wäre interessant, an dieser Stelle weiter zu untersuchen,

inwieweit diese Strategien konfligieren können, beispielsweise weil zwei Nutzer gegenseitig auf die Preisgabe einer Information warten. Außerdem können solche Strategien selbst privat sein, das heißt, dass sie geschützt werden müssen. Es wäre zu analysieren, ob es eine strategische Definition von Strategien erlaubt, die Strategien anderer Nutzer zu erfahren.

Unternehmensinterne Identifikation von Datenschutzverstößen Unser Ansatz zur kollaborativen Identifikation von Datenschutzverstößen erlaubt aktuell die externe Erkennung von Verstößen. Es gibt jedoch auch wesentliche Verstöße, die nur von intern erkannt werden können. Es wäre interessant, die Taxonomie in dieser Hinsicht zu erweitern. Gemäß den gemachten Erfahrungen stellt dies sowohl eine Herausforderung an die Informatik, die die teils komplexen Datenflüsse nachvollziehen können muss, als auch an die Rechtswissenschaften, die die vorliegende Praxis innerhalb eines Unternehmens in den rechtlichen Kontext einordnen und mögliche Verstöße spezifizieren muss, dar.

(k,m)-Anonymität und t-Closeness Bei der Anonymisierung von Suchprotokollen haben wir die Anonymisierung, wie beschrieben und begründet, mittels des Konzeptes der mengenwertigen Anonymisierung '(k,m)-Anonymität' vorgenommen. Diese birgt jedoch das Homogenitätsproblem, wie für k-Anonymität in Kapitel 2.4.2.1 beschrieben. Es wäre interessant zu untersuchen, welchen Effekt die Kombination der mengenwertigen Anonymisierung mit t-Closeness auf den Nutzen hätte [LLV] und wie ein Algorithmus ausgestaltet sein müsste, der solch eine Anonymität generieren kann.

Anhang

A. Realisierung der CSE-Anwendung

Im Folgenden stellen wir die Umsetzung unserer kollaborativen Suchmaschine (CSE) vor. Sie basiert auf Google, dem Suchmaschinenanbieter mit dem größten Marktanteil in Deutschland. Mit Hilfe unserer CSE können Nutzer gemeinsame Suchprojekte koordinieren, von dem Suchverhalten anderer Nutzer lernen und Freunde über ihre Internetaktivitäten auf dem Laufenden halten.

Heutige CSE bestehen aus drei Hauptkomponenten: Die erste Komponente ist eine klassische Suchmaschine, an die die Nutzer ihre Anfragen absetzen. Die zweite Komponente erlaubt es den Nutzern, mittels der CSE Suchanfragen und interessante Links auszutauschen. Als dritte Komponente unterstützen CSEs die Kommunikation zwischen den kollaborierenden Nutzern. Unsere CSE verfügt über eben diese drei Komponenten. Außerdem hat unsere CSE eine Schnittstelle, über die die Nutzer ihre Privatheitspräferenz definieren können.

A.1. Google Schnittstelle

Suchmaschine Für individuelle Suchen bietet unsere CSE die Standardfunktionalität von Google in dem gewohnten Format von Google an. Das heißt, die Nutzer spüren keinen Unterschied zwischen Google und unserem CSE-Mashup.

Ähnliche Anfragen Unsere CSE bietet den Nutzern die Möglichkeit, ähnliche Anfragen anderer Nutzer einzusehen. Eine Suche nach ‘oracle’ liefert somit die üblichen Ergebnisse von Google, kombiniert mit einer Liste ähnlicher Anfragen, zurück (gelbe Liste in Abbildung A.1). Ein Klick auf den Pfeil hinter einer Anfrage öffnet eine Liste von Links, die andere Nutzer aus dem Suchergebnis zu der dargestellten Anfrage angeklickt haben. Für Firefox Nutzer bieten wir außerdem eine Integration unserer CSE in die Schnellsuchleiste an (Abbildung A.2).

A.2. Austausch von Anfragen und Links

Zusätzlich zu den ähnlichen Anfragen, die unsere CSE automatisch anzeigt, können Nutzer auch in den Suchhistorien anderer Nutzer stöbern. Dies hilft bei gemeinsamen Suchprojekten oder bei Informationsbedürfnissen, die auf alten, schon beantworteten

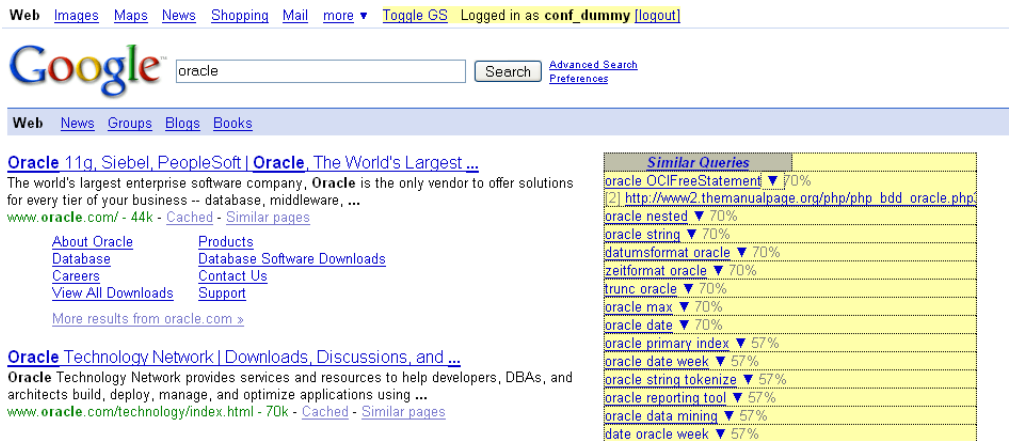


Abbildung A.1.: Anzeige ähnlicher Anfragen und Links

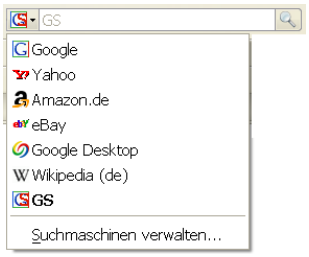


Abbildung A.2.: CSE Integration in den Firefox

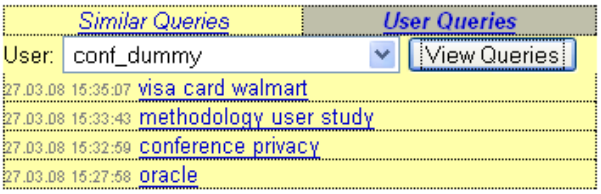


Abbildung A.3.: Durchstöbern der Suchhistorie anderer Nutzer

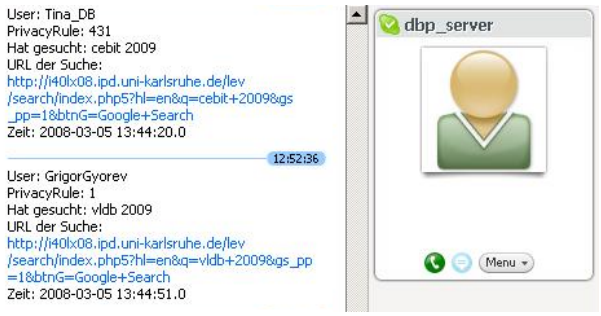


Abbildung A.4.: Skype Benutzerschnittstelle

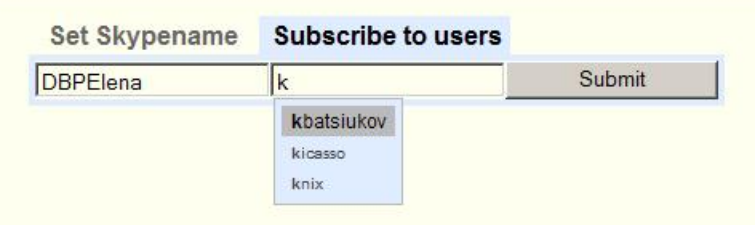


Abbildung A.5.: Abonnements von Suchanfragen und Links anderer Nutzer

Informationsbedürfnissen aufsetzen. Abbildung A.3 zeigt, wie unsere CSE die Anfragen anderer Nutzer darstellt.

Außerdem bietet die CSE eine Skype-Schnittstelle, über die Nutzer Anfragen und angeklickte Links anderer Nutzer abonnieren können (Abbildung A.4). In Skype wird der Nutzername, die Anfrage, die URL und ein Zeitstempel angezeigt. Skype macht den Austausch der Informationen persistent. Abbildung A.5 zeigt, wie Nutzer ein Abonnement erstellen können.

A.3. Kommunikation

Ein wichtiger Aspekt bei CSEs ist die Kommunikation. Skype unterstützt dabei die Funktionen ‘asynchrone Kommunikation’ durch das Cachen von Nachrichten während andere Nutzer offline sind und ’Persistenz’ der Kommunikation durch die integrierte Historienfunktion.

A.4. Schnittstelle zur Definition der Privatheitspräferenz

Es gibt eine Vielzahl von Privatheitsproblemen, wenn Nutzer gegenseitig ihre Suchhistorien einsehen können. Für unsere CSE bieten wir eine Schnittstelle an, über die

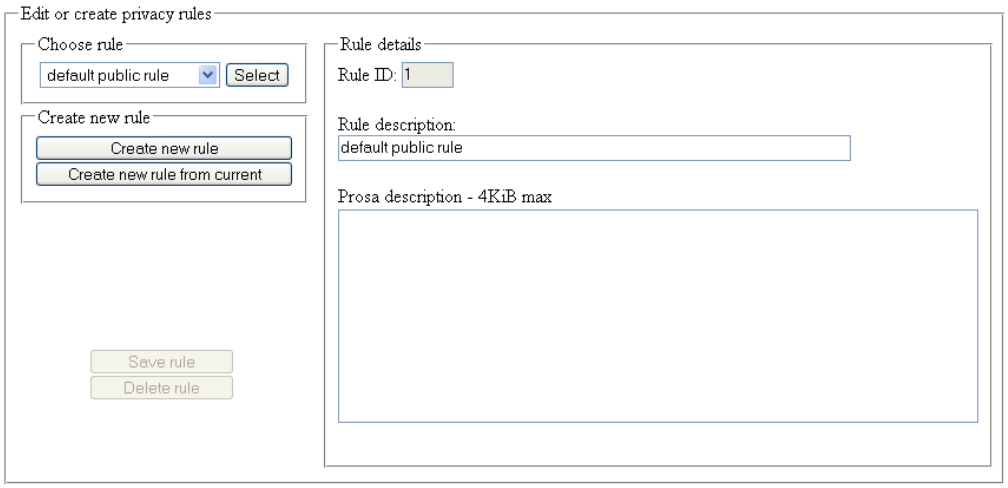


Abbildung A.6.: Strategieeditor für Klartext-Strategien

Nutzer ihre individuelle Privatheitspräferenz definieren können. Unsere CSE bietet die Möglichkeit, die Präferenzen in maschinenlesbarem Format (SQL) (Abbildung A.7) oder in Klartext (Abbildung A.6) zu spezifizieren. Die SQL-Definition kann direkt gegen die zugrundeliegende Datenbank getestet werden. Aus in Kapitel 4.1.2 beschriebenen Gründen, haben wir uns in der Studie rein auf die Klartextstrategien konzentriert. Die Auswahl einer Präferenz (Strategie) erfolgt vor dem Absetzen einer Anfrage (Abbildung A.8). Eine ausgewählte Strategie bleibt aktiv, bis der Nutzer eine andere Strategie auswählt.

A.5. Datenbankschema

Teil der Methodik unserer CSE-Studie ist es gewesen, dass die Teilnehmer umfassende Einblicke in die Implementierung haben. Das unserer CSE zugrundeliegende Datenbankschema stellt Abbildung A.9 dar.

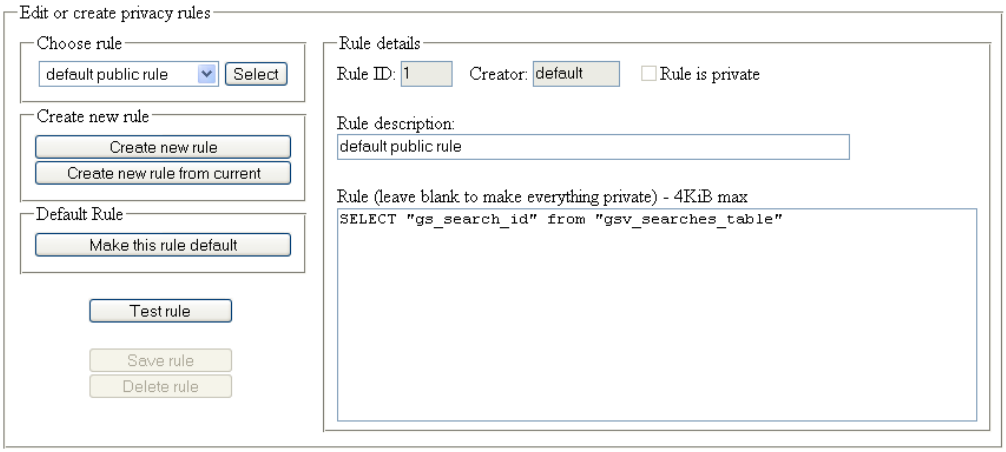


Abbildung A.7.: Strategieeditor für SQL-Strategien

Abbildung A.8.: Auswahl einer Strategie

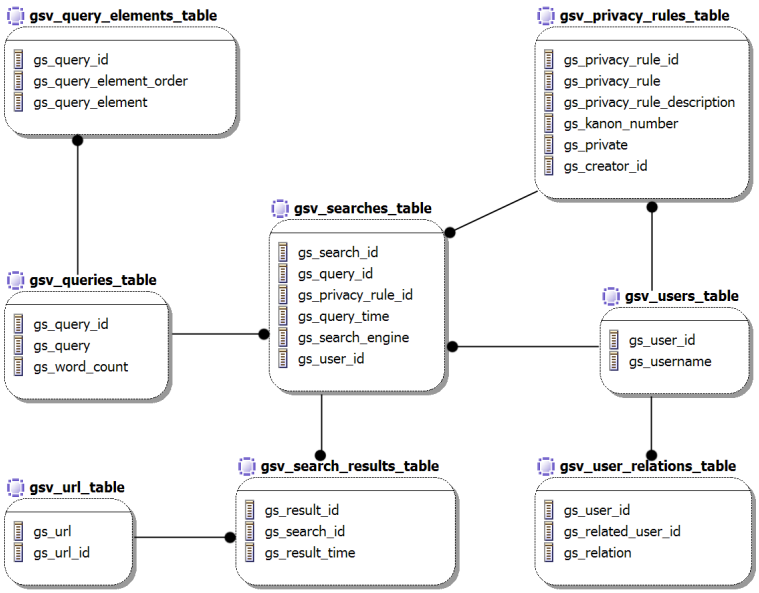


Abbildung A.9.: CSE-Datenbankschema

B. Realisierung der LBS-Anwendung

Im Folgenden geben wir Einblicke in die Umsetzung unserer standortbezogenen Anwendung zum Taggen von Orten. An einigen wenigen Stellen haben wir aus Datenschutzgründen angezeigte Informationen ausgeblendet. Eine detaillierte Beschreibung der einzelnen PETs befindet sich in Kapitel 4.

Im ersten Teil dieses Anhangs beschreiben wir die Web-Applikation (Abschnitt B.1), im zweiten Teil die mobile Anwendung auf dem XDA (Abschnitt B.2). Entsprechend der Beschreibung in Kapitel 4.2 stellen wir die implementierten PETs PET_{fine} und PET_{areas} im Kontext der Web-Applikation und $PET_{checkbox}$, PET_{switch} sowie PET_{anon} im Kontext der mobilen Anwendung vor.

B.1. Web-Applikation

Abbildung B.1 zeigt die Startseite unserer Web-Applikation. Die Startseite ist öffentlich zugänglich, das heißt, jeder Besucher der Seite kann die angezeigten Informationen einsehen. Außerdem kann er Orte suchen (2) und öffentliche Tags durchstöbern. Die Auswahl einer annotierten (getaggten) Position (3) zeigt – soweit diese Informationen nicht geschützt sind – die freigegebenen Tags, die Geo-Koordinaten, den Ersteller, dessen Pseudonym, das Datum und die Uhrzeit an.

Registrierte Nutzer können sich außerdem über diese Seite anmelden (1). Angemeldete Nutzer sehen drei Dialoge: (i) die Karte, (ii) eine Schnittstelle, um Personengruppen zu definieren, auf die Nutzer sich in ihren Privatheitspräferenzen beziehen können, und (iii) einen Dialog für PET_{fine}, mit dem Nutzer Tags für die unterschiedlichen Personengruppen sichtbar machen können. Im Gegensatz zu der öffentlichen Sicht sieht ein angemeldeter Nutzer, welche Tags von ihm sind und welche Tags andere Nutzer für ihn sichtbar gemacht haben. Die dritte Kategorie von Tags sind Informationen, die wir aus Wikipedia extrahiert haben. Mit (Abbildung B.2, (4)) kann der Nutzer festlegen, welche Annotationen angezeigt werden sollen. Außerdem hat der Nutzer (5) kontinuierlich Einblick, wie viele Frei-SMS er sich verdient hat, einmal aufgrund der Zeit, die er online gewesen ist, und einmal für getaggte Orte. Mit Hilfe des Dialogs in Abbildung B.3 definiert ein Nutzer für jede Personengruppe deren reale Repräsentanten (6). Diese werden dann in (7) angezeigt, also beispielsweise in ‘Eltern’ ‘Mutter von X’ etc.

PET_{fine} dient der feingranularen Definition von Privatheitspräferenzen, hier im ersten Schritt für Tags. Abbildung B.4 zeigt, wie Nutzer Tags (8) und die unterschiedlichen

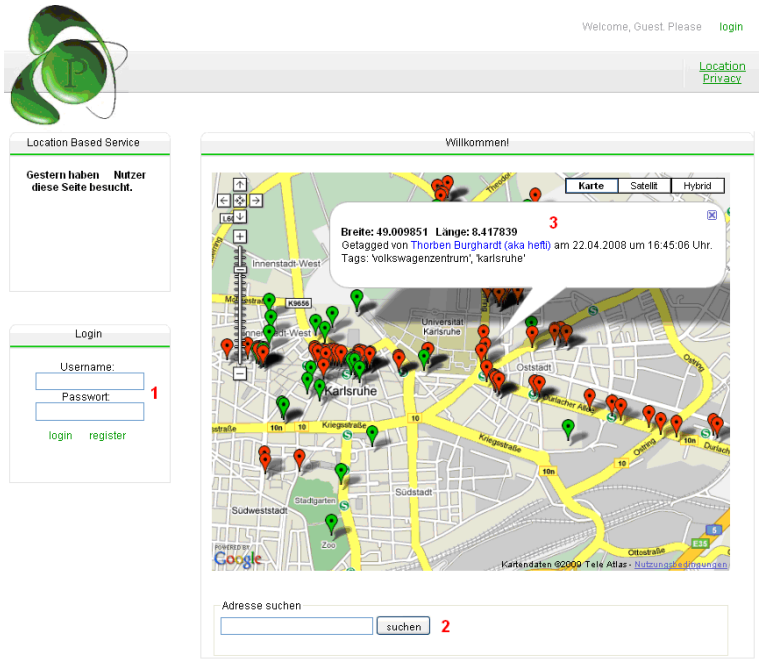


Abbildung B.1.: Startseite und Übersicht

Metadaten (9) den verschiedenen sozialen Gruppen (10) preisgeben können. Eine Karte (linker Teil der Abbildung) gibt die Position des gerade betrachteten Tags an.

Das zweite in der Web-Applikation implementierte PET ist PET_{areas} (Abbildung B.5). Hier können die Nutzer mehrere Bereiche definieren, innerhalb derer sie vor jedem oder bestimmten Personengruppen geschützt sein möchten (11). Diese Bereiche kann der Nutzer auf der Karte in Form (geschlossener) Polygone spezifizieren (12). Der eingezeichnete Bereich entspricht beispielsweise dem Arbeitsumfeld eines Informatikers am KIT. Wie zwischen den unterschiedlichen Personengruppen unterschieden werden kann, verdeutlicht Abbildung B.6. Den Effekt eines geschützten Bereiches spiegelt Abbildung B.7 wider. Zu jedem Kalendertag werden alle möglichen Tracks angezeigt (13) und wahlweise auf der Karte eingeblendet (14). Betritt oder verlässt ein Nutzer einen seiner geschützten Bereiche (15), so wird das GPS deaktiviert. Das heißt, der Track ist für die Gruppen, vor denen der Bereich den Nutzer schützen soll, nicht mehr sichtbar.

Zuletzt haben wir PET_{fine} so erweitert, dass es sich auch auf Tracks anwenden lässt (Abbildung B.8). Wie bei der Anwendung auf Tags kann der Nutzer einen Track und die damit einhergehenden Metadaten den unterschiedlichen sozialen Gruppen preisgeben. Auf einer Karte zeigt unsere Anwendung den jeweiligen Track an.

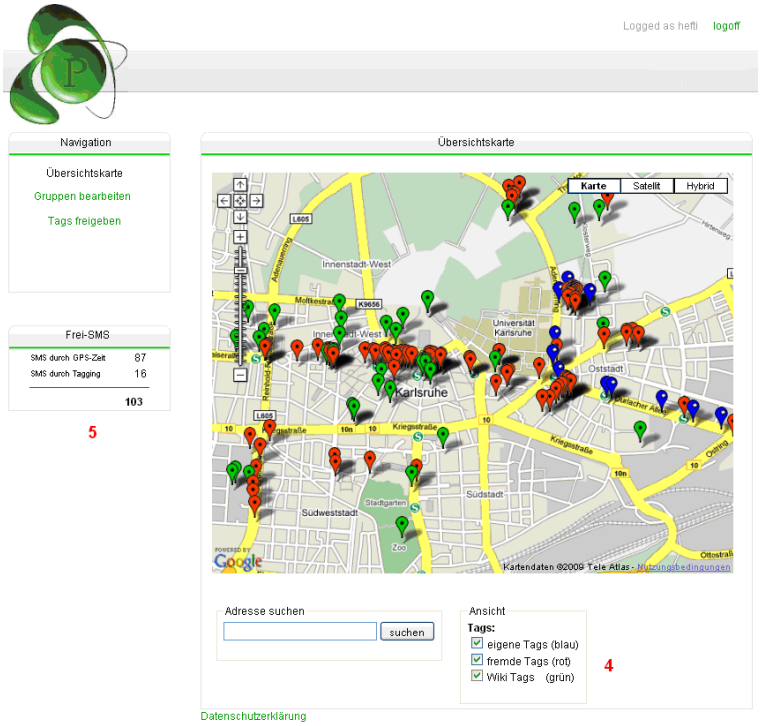


Abbildung B.2.: Private Ansicht der Web-Applikation

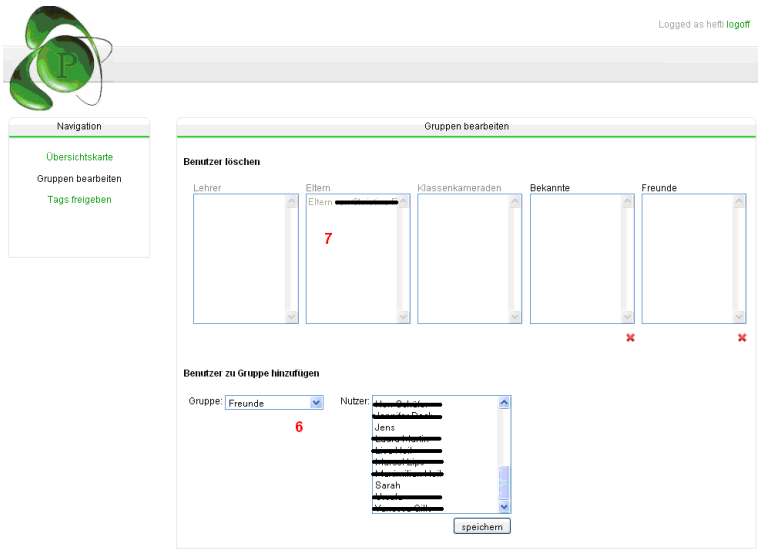


Abbildung B.3.: Definition der Personengruppen

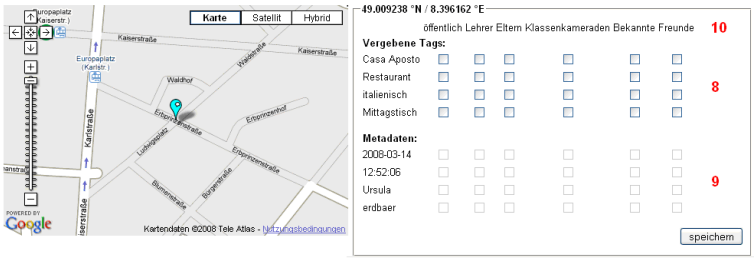


Abbildung B.4.: Realisierung von PET_{fine}

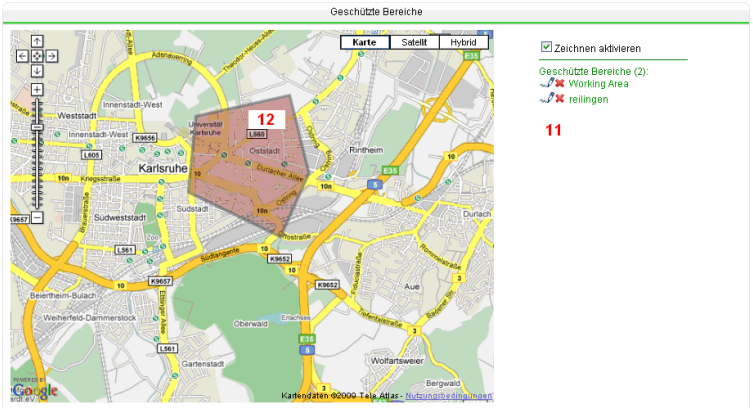


Abbildung B.5.: Realisierung von PET_{areas}

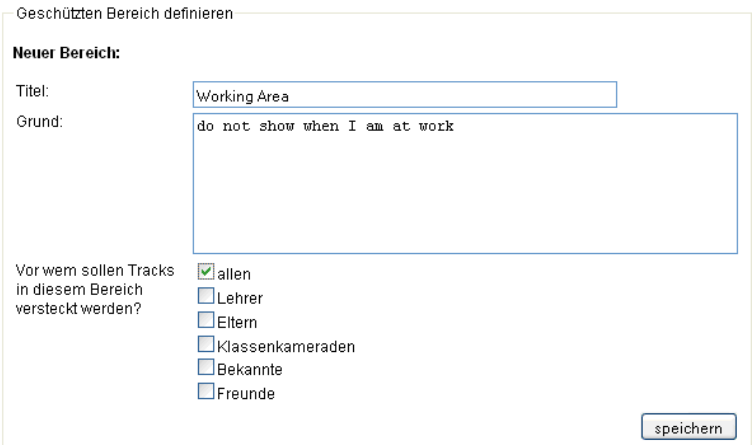


Abbildung B.6.: Definition geschützter Bereiche

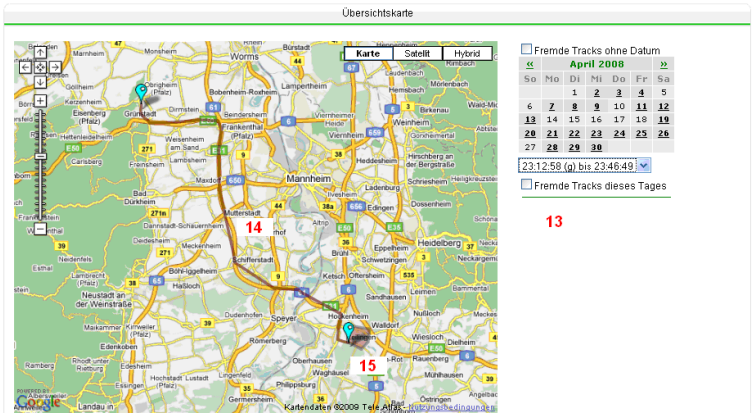


Abbildung B.7.: Geschützte Bereiche im Einsatz

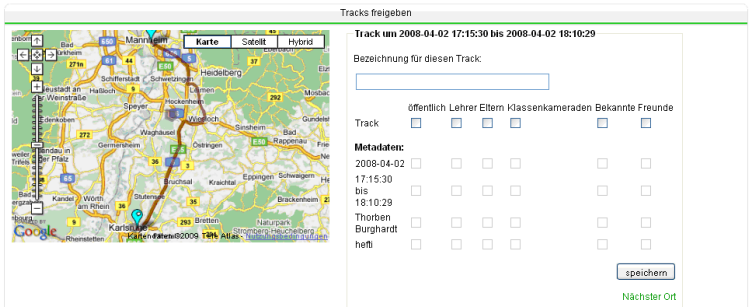


Abbildung B.8.: Realisierung von PET_{fine} für Tracks

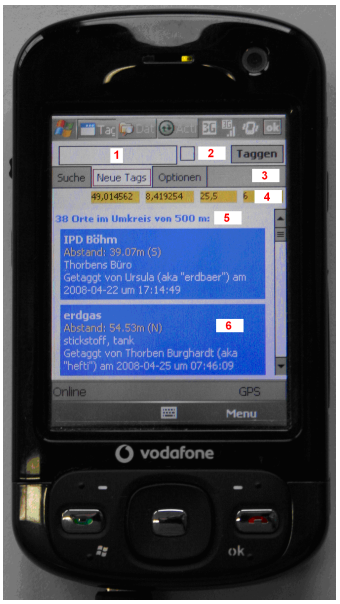


Abbildung B.9.: $PET_{checkbox}$

Abbildung B.10.: PET_{switch}

Abbildung B.11.: PET_{anon}

B.2. Mobile Anwendung auf dem XDA

Im Alltag begleitet den Nutzer unsere mobile Anwendung auf dem XDA. Das XDA sendet dabei die aktuelle Position an unser System und bekommt daraufhin relevante Informationen zum aktuellen Umfeld zurückgeschickt. Diese zeigt die mobile Anwendung dann dem Nutzer an (Abbildung B.9). (4) zeigt die Position des Nutzers und die Genauigkeit, mit der das GPS diese Position aktuell erkennen kann. (5) weist aus, in welchem Umkreis nach relevanter Information gesucht wird und wie viele Informationen aktuell vorliegen. (6) ist ein Beispiel für eine annotierte Lokation. Dazu gibt unsere Anwendung an, wie weit der getaggte Ort vom aktuellen Standort des Nutzers entfernt ist, wer den Ort wann annotiert hat und in welche Himmelsrichtung der Ort liegt. Die Metadaten zu der Erstellung des Tags und die Tags selbst sind nur verfügbar, wenn der Ersteller sie preisgibt. Abbildung B.10 zeigt die Integration von PET_{switch}. Über die Taste (7) kann der Nutzer auf einfache Art und Weise das GPS an- beziehungsweise abschalten. Natürlich kann unsere Anwendung dann auch die Informationen im Umfeld des aktuellen Standortes nicht mehr anzeigen. Bei der Nutzung von PET_{anon} (Abbildung B.11) wählen die Nutzer ein k, das definiert, von wie vielen Personen sie ununterscheidbar sein möchten. Wann immer die Anzeige mit den Informationen im aktuellen Umfeld des Nutzers aktualisiert wird, zeigt PET_{anon} während der Berechnung die Anzahl gefundener weiterer Nutzer an (8). Je größer k, desto länger dauert die Berechnung, bis aktualisierte Umfeldinformationen angezeigt werden können.

C. Anbieter Vollzugsdefizitanalyse

Online-Shops
Amazon
Otto
Quelle
Deutsche Bahn
Neckermann
Tchibo
Expedia
Hotel Reservation
Weltbild
Karstadt
Vodafone
Autoteile
Autoverwertung Wieben
Renet
Buch.de
Allago
Officio
Dell.com
Softline
Softwarehouse
EP Netshop
Hifi Edition
Beautynet
Schlecker
Cocktailbude
Gourmondo
MyMuesli
AllNatura
SM-Küchen
Apple Store
bol.de
JPC
Musicload.de
Douglas
Avon
Reifen Direkt
Reifen Fix
Deichmann
Goertz
MyToys
Kanu-Gatz
Sportcheck
Herrenausstatter
Service Shirtcity
Yoox
Deutsche Internetapotheke
Mycare
Sanicare

Informationsportale
Spiegel Online
Chefkoch
billiger.de
FOCUS Online
Welt.de
wetter.com
sueddeutsche.de
Falk.de
WetterOnline
goFeminin.de
FAZ.NET
computerbild.de
preisvergleich.de
Verivox.de
Ciao.de
Heise.de
ProSieben
Sat1
Die ZEIT
RTL
Swr3Land

E-Mail-Dienste
Abacho
AOL
Arcor
Freenet
GMX
GoogleMail
Lycos
Microsoft
Unicum
Web.de

Soziale Netzwerks.
StudiVZ
MySpace
Wer-Kennt-Wen
Lokalisten
Xing
Jappy
StayFriends

Suchmaschinen
Google
Yahoo!
MSN
AOL
Yasni
123People

Auktionshäuser
Ebay
Auvito.de
hood.de
www.besteauktion.de
www.compendo.de
www.azubo.de
www.auxion.de
www.amprice.de

D. Realisierung von *CLEF* und der Taxonomie

In diesem Anhang geben wir weiterführende Informationen zu Kapitel 5.2. Abschnitt D.1 gibt Einblicke in die Umsetzung unseres Web 2.0-Ansatzes. Abschnitt D.2 beschreibt die Taxonomie, das heißt die Abhängigkeiten der Fragen. Abschnitt D.3 gibt jede Frage wie in *CLEF* dargestellt wieder, Abschnitt D.4 beschreibt die daraus resultierenden Datenschutzverstöße.

D.1. *CLEF* Realisierung

D.1.1. Startseite und Registrierung

Abbildung D.1 zeigt die Startseite und gleichzeitig das Registrierungsformular von *CLEF*. Wir unterscheiden die Rollen 'Internetnutzer' (Inspektoren), 'Unternehmen' und 'Datenschutzbehörden'. Zur Registrierung muss ein Nutzer seinen Namen und ein Passwort hinterlegen. Möchte der Nutzer kontinuierlich über identifizierte Datenschutzverstöße auf dem Laufenden gehalten werden, so kann er (optional) eine E-Mailadresse hinterlegen.

D.1.2. Anbieterübersicht

Nach der Anmeldung können die Nutzer entweder selbst einen Anbieter auswählen, den sie überprüfen möchten, oder *CLEF* einen Anbieter vorschlagen lassen. In der beschriebenen Studie hat *CLEF* die Anbieter automatisch vorgeschlagen. Abbildung D.2 zeigt die Übersichtsseite direkt nach der Anmeldung. In der ersten Spalte zeigen wir den Anbieter an, in der dritten Spalte die beantworteten Fragen unserer Taxonomie und die Anzahl offener Fragen pro Anbieter. Bei der Studie haben die Teilnehmer für jede Frage angegeben, ob sie diese *einfach, mittel* oder *schwierig* zu beantworten fanden. Spalte vier zeigt an, ob in jedem Fall auch der Schwierigkeitsgrad beantwortet wurde. Zum Starten der Überprüfung eines Anbieters muss der Nutzer 'bearbeiten' (in der zweiten Spalte) auswählen.

Ein Nutzer der Rolle 'Datenschutzaufsichtsbehörde' hat eine andere Sicht als der Inspektor (Abbildung D.4). Mitarbeiter der Behörden können (8) die Anzahl der 'Ja'-

Abbildung D.1.: Startseite und Registrierung

Anbieterübersicht

1 Anbieter:	2 Bearbeiten:	3 Beantwortete Fragen:	4 Schwierigkeit bewertet:
Anonymized	Bearbeiten	0 (9 offen)	0 (9 offen)
Anonymized	Bearbeiten	0 (9 offen)	0 (9 offen)
Anonymized	Bearbeiten	0 (9 offen)	0 (9 offen)
Anonymized	Bearbeiten	0 (9 offen)	0 (9 offen)
Anonymized	Bearbeiten	0 (9 offen)	0 (9 offen)
Anonymized	Bearbeiten	0 (9 offen)	0 (9 offen)

Abbildung D.2.: Initiale Anbieterübersicht

Anbieterübersicht

Anbieter:	Bearbeiten:	Beantwortete Fragen:	Schwierigkeit bewertet:	Identifizierte Verstöße:
Anonymized	Bearbeiten	2 (14 offen)	2 (14 offen)	0
Anonymized	Bearbeiten	28	28	3 6 Ausblenden
Einzelne Verstöße: 7 - Kein Hinweis, wie lange Daten gespeichert werden (§ 13 Abs. 1 S. 1 TMG) - Falsche Angabe der Cookie-Speicherdauer. (§ 13 Abs. 1 S. 2 TMG) - Unterrichtung über automatisierte Verfahren finden nicht zu Beginn statt. (§ 13 Abs. 1 S. 2 TMG)				
Anonymized	Bearbeiten	28 (7 offen)	28 (7 offen)	5 Anzeigen

Abbildung D.3.: Anbieterübersicht und Verstöße

Frage 08:	8 Ja: 91% (11)	9 Nein: 8% (1)	10 Ja	Ändern
Frage 09:	Ja: 100% (12)	Nein: 0% (0)	Ja	Ändern
Frage 10:	Ja: 100% (12)	Nein: 0% (0)	Ja	Ändern
Frage 11:	Ja: 91% (11)	Nein: 8% (1)	Ja	Ändern
Frage 23:	Ja: 58% (7)	Nein: 41% (5)	11 Nein	Ändern

Abbildung D.4.: Grad der Übereinstimmung (Behördenperspektive)

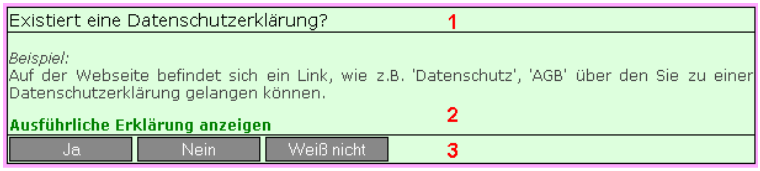


Abbildung D.5.: Beispielfrage der Taxonomie Frage

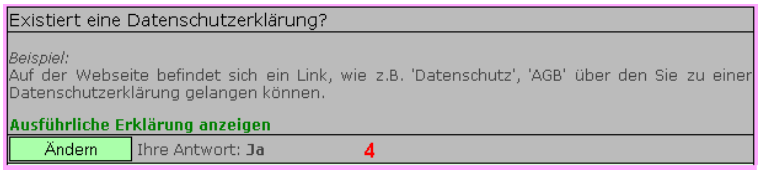


Abbildung D.6.: Beantwortete Frage aus der Taxonomie

und (9) die Anzahl der 'Nein'-Antworten, aggregiert über alle Inspektoren, sehen. Außerdem sehen sie (10) ihre eigene Antwort mit einer roten Hervorhebung (11), wenn die eigene Antwort nicht mit der Gemeinschaftsantwort der Inspektoren übereinstimmt. Außerdem fließen die Antworten von Behörden nicht in die Gemeinschaftsantwort mit ein. Nochmal der Hinweis, dass die Teilnehmer unserer Studie 'Inspektoren' gewesen sind, sie konnten diese Sicht also nicht sehen.

Ein Klick auf 'Einblenden' in der fünften Spalte (Abbildung D.3) öffnet für den ausgewählten Anbieter eine Liste identifizierter Verstöße. Außerdem zeigt es die relevante rechtliche Grundlage für den Verstoß an.

D.1.3. Implementierung der Taxonomie

Die Taxonomie besteht aus mehreren Ausgangsfragen auf der untersten Ebene und Fragen, die von diesen Ausgangsfragen abhängig sind.

Abbildung D.5 zeigt eine einfache Ausgangsfrage, die nach der Existenz einer Datenschutzerklärung fragt. In der ersten Zeile (1) steht die eigentliche Frage, in der Mitte (2) ein Kurzbeispiel und darunter (3) drei Bedienelemente zur Beantwortung der Frage mit ´Ja', ´Nein', 'Weiß nicht'. Eine beantwortete Frage wird, wie in Abbildung D.6 dargestellt, ausgegraut. Über 'Ändern' kann die Antwort geändert werden (4).

Durch die Auswahl von 'Ausführliches Beispiel anzeigen' kann ein Nutzer eine detaillierte Beschreibung zu jeder Frage einblenden (Abbildung D.7, (5)). Wir geben in dieser Beschreibung Informationen über den rechtlichen Hintergrund, das Ziel des Gesetzgebers etc. an. Die detaillierten Beschreibungen wurden dabei insbesondere von Mitarbeitern der Arbeitsgruppe von Prof. Kühling erstellt.

Immer, wenn ein Nutzer eine Frage beantwortet oder seine Antwort ändert, fragen

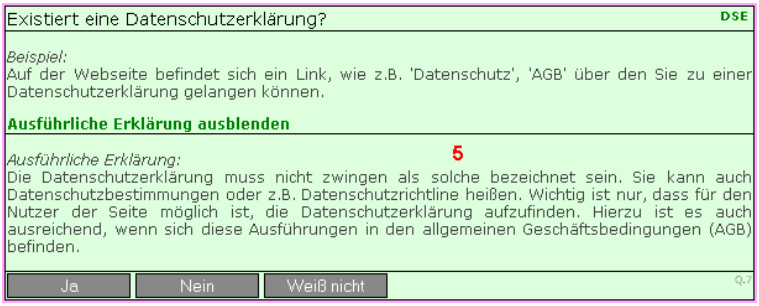


Abbildung D.7.: Frage mit ausführlicher Erklärung

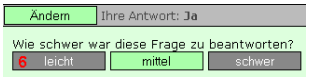


Abbildung D.8.: Bewertung der Schwierigkeit einer Frage

wir ihn nach der Schwierigkeit der Frage (Abbildung D.8, (6)) .

D.1.4. Informationsquellen

Zur Beantwortung der meisten Fragen muss der Nutzer die Datenschutzerklärung durchsuchen, sich bei dem Anbieter registrieren, die gespeicherten Cookies untersuchen etc. Wir haben einen Crawler integriert, der automatisch versucht, den Link auf die allgemeinen Geschäftsbedingungen, die Datenschutzerklärung, das Impressum und das Registrierungsformular zu finden. Außerdem erkennt unser Crawler Webstatistikwerkzeuge und Cookies, die der betrachtete Anbieter einsetzt. Für jede Frage bietet *CLEF* Links an, die direkt auf die Informationsquellen verweisen, die wir für die Beantwortung einer Frage für relevant halten (rechte obere Ecke von Abbildung D.9, (7)). Eine besondere Informationsquelle stellen die 'Live Daten' dar, das heißt die Informationen, die wir automatisiert erheben (Abbildung D.10). In unserem Fall ist das die Erkennung der eingesetzten Cookies und der Webstatistikwerkzeuge.

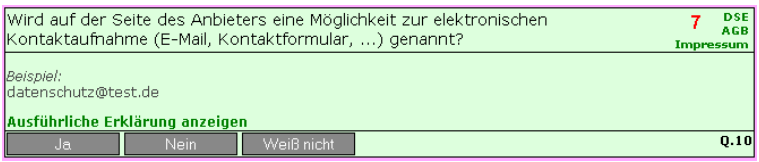


Abbildung D.9.: Frage mit drei Informationsquellen

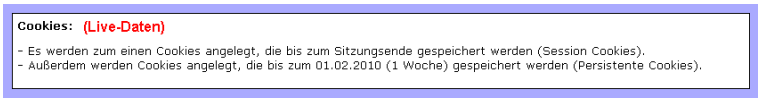


Abbildung D.10.: Live-Daten für Cookies

D.2. Struktur der Taxonomie

Das Ziel dieses Abschnittes ist es, die Struktur und die Fragen der in Kapitel 5.2 entworfenen Taxonomie darzustellen. Die Taxonomie kann auf unterschiedliche Weise modelliert werden. Wir haben gemäß Erfahrungen aus unseren Vorarbeiten (Kapitel 3) die Taxonomie so konstruiert, dass Nutzer die Fragen, entsprechend des Aufbaus der Datenschutzerklärungen vieler Anbieter, möglichst eine nach der anderen beantworten können.

Nomenklatur Wir stellen die Taxonomie in Form von (Teil-) Graphen dar. Es gibt fünf Arten von Knoten:

Themenknoten Rechteckige Knoten mit gestrichelten Linien fassen Details, zugehörig zu dem jeweils gleichen Thema, zusammen. Jeder Themenknoten wird auch detailliert dargestellt.

Frageknoten Jede Frage wird durch einen rechteckigen Knoten mit durchgezogener Linie repräsentiert. Kurzformen der Fragen sind innerhalb des Knotens angegeben.

Verstoßknoten Verstöße sind als graue Rechtecke modelliert. Pro dargestelltem Graph fassen wir alle möglichen Verstöße in einem Knoten zusammen. Eine eingehende Kante beschreibt, aus welcher Frage ein Verstoß resultiert.

Finalknoten Wann immer es keine weiteren Fragen zu einem Thema zu beantworten gibt, verweist die Antwort auf die letzte Frage auf den Finalknoten.

Operatorknoten Es kann mehrere vorangegangene Antworten geben, die zu der Beantwortung der gleichen Folgefrage führen. Zur besseren Darstellung geben wir jede Frage nur einmal an und verknüpfen vorangegangene Antworten mit diamantförmigen `oder`- und `und`-Knoten.

Die Kanten zwischen Knoten beschreiben jeweils, nach welcher Antwort *CLEF* welche Frage stellt. Die ja / nein Antworten sind an den Kanten annotiert.

Taxonomieübersicht Die Taxonomie ist unterteilt in unterschiedliche Themenbereiche (Abbildung D.11). Zuerst wird geklärt, ob das TMG, als der hier betrachtete Rechtstrahmen, angewendet werden kann (wir haben für unsere Studie nur solche Anbieter ausgesucht, die dieses Kriterium auch erfüllen). Anschließend kann *CLEF* konkretere Fragen zu der Datenschutzerklärung, der Datenerhebung und der Einwilligung stellen.

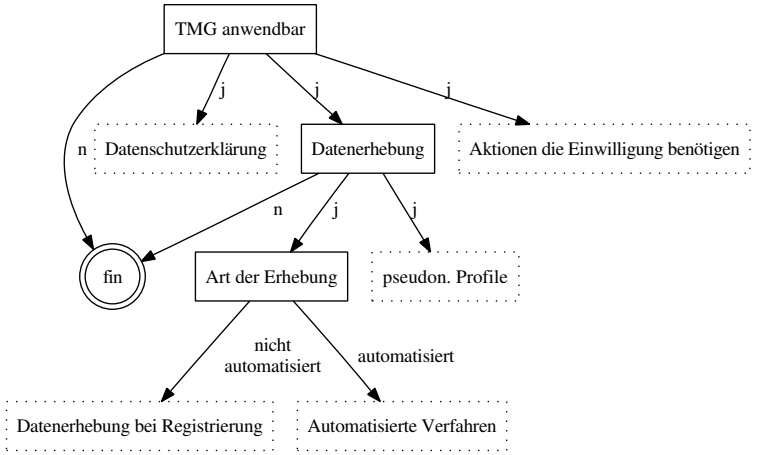


Abbildung D.11.: Übersicht der Taxonomie

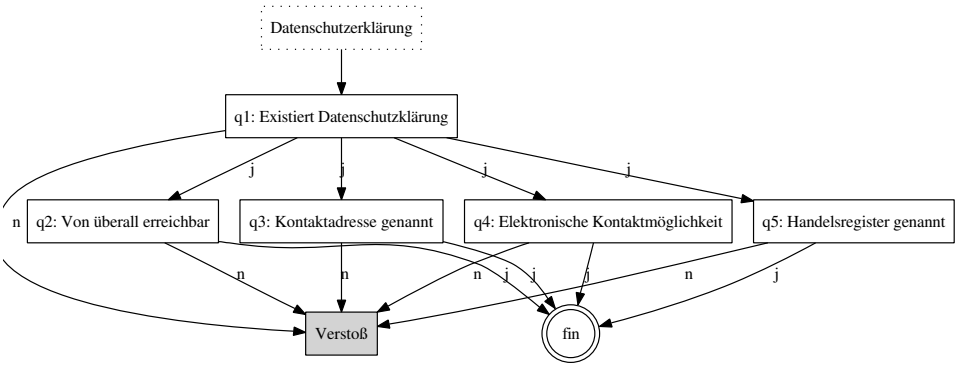


Abbildung D.12.: Taxonomieauszug: Datenschutzerklärung

Die Datenerhebung ist weiter unterteilt in die Datenerhebung bei der Registrierung und der Erhebung durch automatisierte Verfahren.

Datenschutzerklärung Das Thema 'Datenschutzerklärung' (Abbildung D.12) umfasst die Prüfung auf die Existenz einer Datenschutzerklärung. Sollte eine Datenschutzerklärung existieren, wird überprüft, ob diese von überall aus erreichbar ist, ob eine Kontaktadresse zu einem Datenschuztverantwortlichen genannt ist, ob der Kontakt elektronisch hergestellt werden kann und ob das Handelsregister genannt wird.

Datenerhebung bei der Registrierung Die detaillierte Betrachtung des Taxonomieauszuges für das Thema 'Datenerhebung bei Registrierung' (Abbildung D.13) prüft,

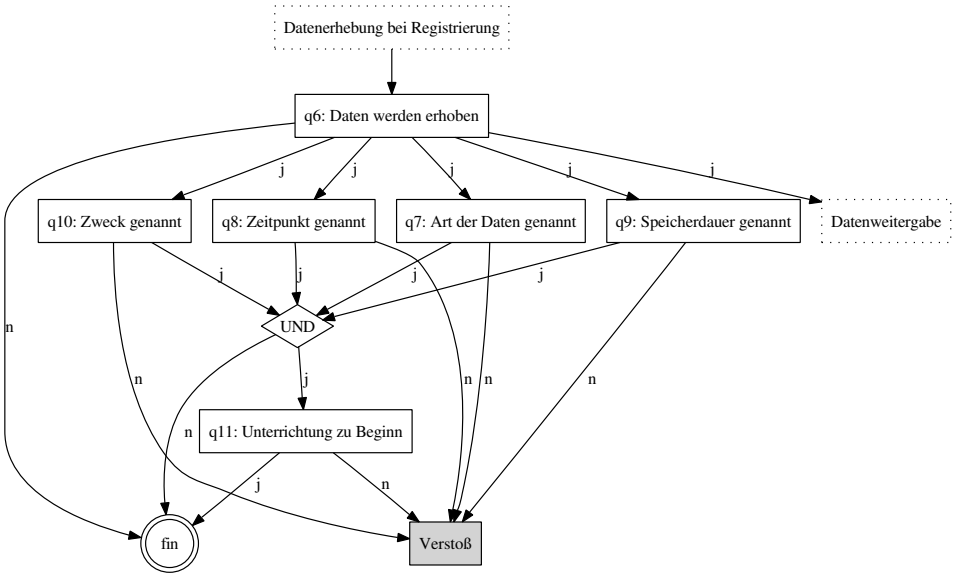


Abbildung D.13.: Taxonomieauszug: Datenerhebung bei der Registrierung

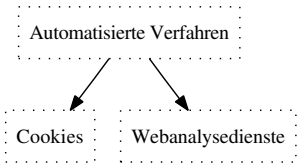


Abbildung D.14.: Übersicht automatisierter Verfahren

ob der Anbieter über Art, Zweck und Umfang (Zeitpunkt der Datenerhebung und Speicherdauer) informiert. Nur wenn er alle diese Informationen ausweist und dies auch vor der Erhebung macht, liegt kein Verstoß vor. Außerdem folgt unter Umständen die Überprüfung, ob eine Datenweitergabe durch den Anbieter erfolgt.

Automatisierte Verfahren Wie Abbildung D.14 zeigt, unterscheiden wir die automatisierten Verfahren in Cookies und Webanalysedienste bzw. Webstatistikwerkzeuge (wir benutzen beide Begriffe gleichbedeutend). Es existieren weitere Verfahren wie Web-Bugs, die wir hier jedoch nicht modellieren.

Die Taxonomie zur Cookie-Nutzung umfasst (Abbildung D.15) die Überprüfung, ob Cookies eingesetzt werden, ob es sich dabei um persistente oder Session-Cookies handelt, ob auf das jeweilige Cookie ein Hinweis erfolgt und ob die Speicherzeit sowie der Zweck des Cookies genannt werden.

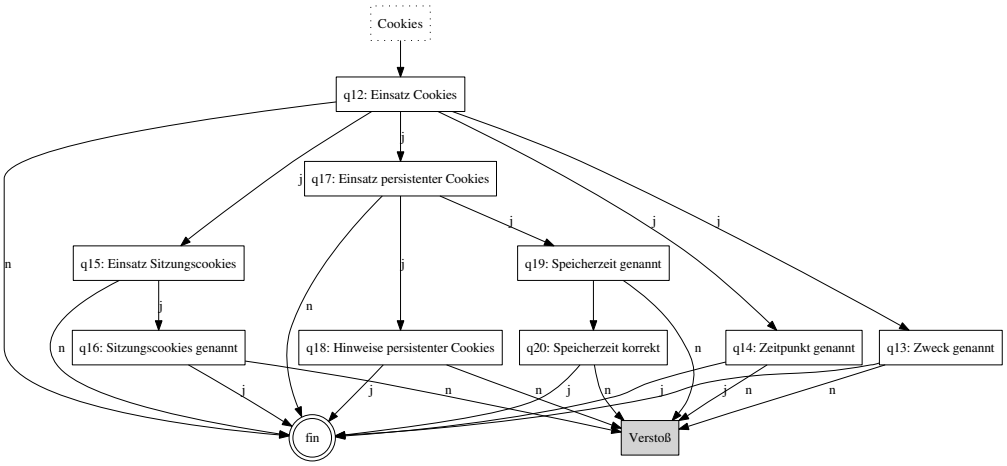


Abbildung D.15.: Taxonomieauszug: Cookies

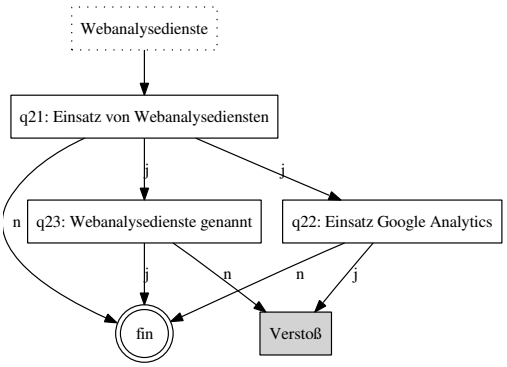


Abbildung D.16.: Taxonomieauszug: Webanalysedienste

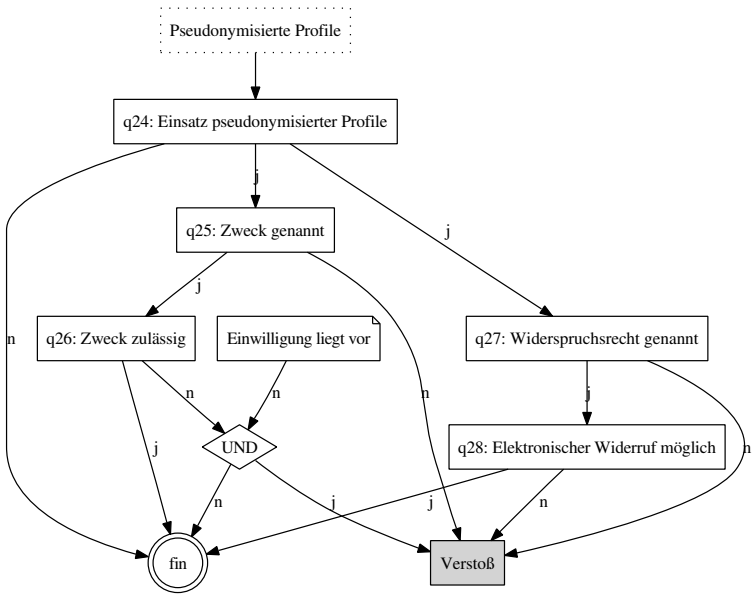


Abbildung D.17.: Taxonomieauszug: Pseudonymisierte Profile

Für Webanalysedienste haben wir mittels der Taxonomie geprüft, ob besagte Dienste eingesetzt werden und ob ein Hinweis darauf erfolgt. Da Google Analytics nicht konform zu geltendem Recht [Sok09, Seite 34], [Obe09] ist, ist dessen Einsatz automatisch ein Verstoß.

Pseudonymisierte Profile Der Telemedienanbieter darf zu von dem Gesetzgeber vorgegebenen Zwecken aus Nutzungsdaten, wie der URL, von der aus ein Nutzer die Seite besucht hat, oder dem Browsertyp, pseudonymisierte Nutzungsprofile erstellen. Wir prüfen mit der Taxonomie, ob der Anbieter den Zweck der Profilerstellung ausweist und gegebenenfalls eine Einwilligung vom Benutzer einholt. Dies ist erforderlich, falls es sich bei dem Zweck nicht um Werbung, Marktforschung oder die bedarfsgerechte Gestaltung der Telemedien handelt. Außerdem prüfen wir, ob der Anbieter den Nutzer auf sein Widerspruchsrecht für die Erstellung pseudonymisierter Profile hinzuweist und ob er ihm die Möglichkeit gibt, diesen Widerspruch ohne Medienbruch durchzuführen. Das geht beispielsweise durch den Klick auf eine Schaltfläche oder über ein Kontaktformular.

Datenweitergabe Abbildung D.18 beschreibt den Auszug unserer Taxonomie zu dem Thema ‘Datenweitergabe’. Befindet sich das Empfängerunternehmen außerhalb der EU und liegt dort kein von der EU anerkanntes, ausreichendes Schutzniveau vor, muss der

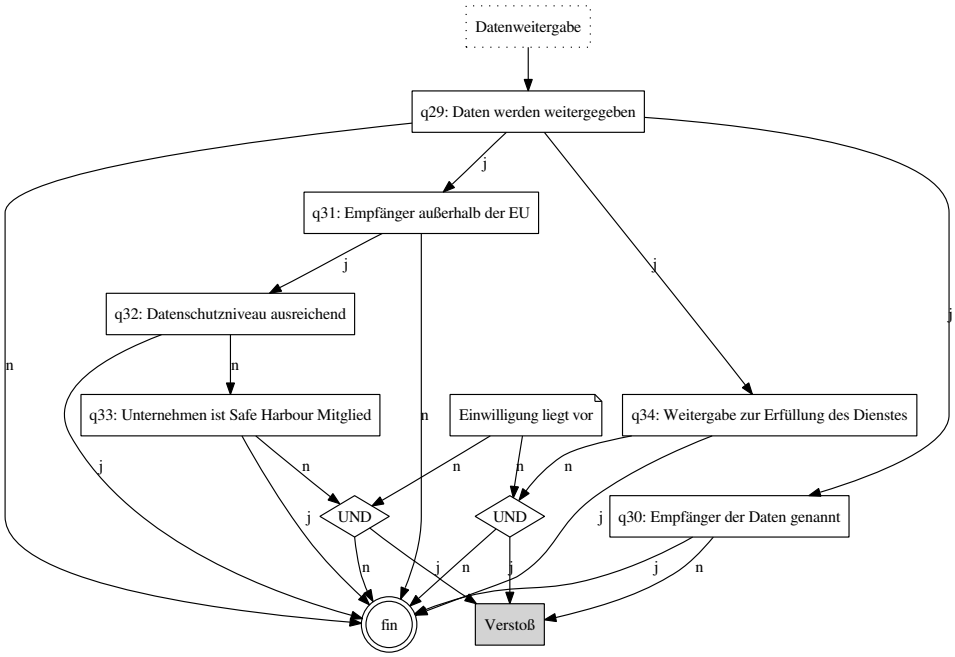


Abbildung D.18.: Taxonomieauszug: Datenweitergabe

Nutzer in die Datenweitergabe einwilligen. Eine Ausnahme stellen die USA dar, so der Empfänger Mitglied des Safe-Harbor Agreements ist. Die Empfänger sind auszuweisen. Abhängig vom Zweck der Weitergabe ist außerdem eine Einwilligung erforderlich.

Einwilligung Neben oben genannten Gründen für eine Einwilligung, gibt es zwei weitere häufige Situationen, die eine Einwilligung erfordern (Abbildung D.19): Die Erstellung personalisierter Profile und die Erhebung von Daten, die über solche Daten hinausgehen, die zur Diensterbringung erforderlich sind. Muss der Nutzer einwilligen, so muss die Einwilligungserklärung jederzeit abrufbar sein, der Anbieter muss das Recht auf Widerruf nennen und die Einwilligung gegebenenfalls graphisch hervorheben. Erfüllt er eine der Anforderungen nicht, liegt ein Verstoß vor. Als dritten Punkt haben wir in einer Studie innerhalb unserer Vorarbeiten (Kapitel 3) beobachtet, dass einige Anbieter eine Einwilligung in unzulässige Forderungen einholen, so zum Beispiel der Verzicht auf ein Widerrufsrecht oder der Verzicht auf das Recht des Auskunftsersuchen.

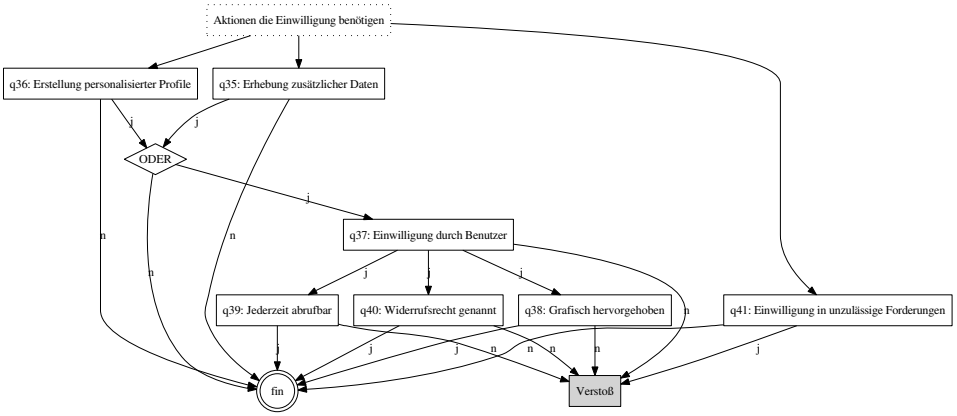


Abbildung D.19.: Taxonomieauszug: Einwilligung

D.3. Fragen der Taxonomie

Im Folgenden stellen wir zu allen Themen die dazugehörigen Fragen vor. Jede Frage besteht aus einem Identifikator (*FragenID*), der Frage selbst, einem kurzen und einem ausführlichen Beispiel. Folgefragen der Ausgangsfragen jeder Kategorie sind eingerückt dargestellt.

FragenID: Fragestellung
Kurzes Beispiel
Ausführliches Beispiel (optional)

Datenschutzerklärung

q1: Existiert eine Datenschutzerklärung?
Auf der Webseite befindet sich ein Link, wie z.B. 'Datenschutz', 'AGB'; über den Sie zu einer Datenschutzerklärung gelangen können.
Die Datenschutzerklärung beschreibt unter anderem, welche persönlichen Daten ein Anbieter sammelt, und wie er mit diesen Daten umgeht. Die Datenschutzerklärung muss nicht zwingend als solche bezeichnet sein. Sie kann auch 'Datenschutzbestimmungen' oder z.B. 'Datenschutzrichtlinie' heißen. Wichtig ist nur, dass es für den Nutzer der Seite möglich ist, die Datenschutzerklärung aufzufinden. Hierzu ist es auch ausreichend, wenn sich diese Ausführungen in den allgemeinen Geschäftsbedingungen (AGB) befinden.

q2: Ist die Datenschutzerklärung von jeder Seite des Anbieters aus zu erreichen?
Existiert sowohl auf der Startseite, als auch auf jeder Unterseite ein Link, über den man zur Datenschutzerklärung gelangen kann?
Natürlich ist von extern (für den Nutzer) nicht wirklich jede Seite überprüfbar. Die Frage kann aber mit *ja*; beantwortet werden, wenn z.B. der Link auf die Datenschutzerklärung Teil des Seitenrahmens ist und dieser auch nach dem Wechsel zwischen zwei oder mehreren Seiten gleich ist. Meistens befindet sich solch ein Link am unteren Rand der Seite oder ganz rechts / links.

q3: Wird auf der Seite (z.B. in der Datenschutzerklärung, in den AGB oder im Impressum) die Kontaktadresse des Anbieters genannt (Name und Anschrift)?
Test AG, Hauptstraße 1, Berlin
Der Anbieter ist verpflichtet, seinen Namen und die Anschrift zu nennen.

q4: Wird auf der Seite des Anbieters eine Möglichkeit zur elektronischen Kontaktaufnahme (E-Mail, Kontaktformular, ...) genannt?
datenschutz@test.de
Dem Nutzer der Seite muss eine Möglichkeit gegeben sein, mit dem Anbieter schnell auf elektronischem Wege Kontakt aufzunehmen (Bsp.: E-Mail-Adresse; Kontaktformular,...).

Wird in der Datenschutzerklärung, in den AGB oder im Impressum das Handelsregister, Vereinsregister, Partnerschaftsregister oder Genossenschaftsregister genannt, in das der Anbieter eingetragen ist?
Handelsregister B Fürth HRB 11927
Der Anbieter der Seite ist verpflichtet darauf hinzuweisen, in welches Register er eingetragen ist. Die Angabe des Registers variiert je nach Rechtsform. (eingetragener Kaufmann (Handelsregister); GmbH (Handelsregister); Verein (Vereinsregister) usw.)

Datenerhebung bei Registrierung

q6: Wird in der Datenschutzerklärung darauf hingewiesen, dass personenbezogene Daten erhoben werden?
Für die Ausgestaltung dieses Vertragsverhältnisses werden Ihre Daten für die Nutzung unserer Dienste erhoben.
An dieser Stelle geht es nur um die grundsätzliche Frage, ob überhaupt darauf hingewiesen wird, dass personenbezogene Daten erhoben werden. Personenbezogene Daten sind alle Angaben, die einer bestimmten Person zugeordnet werden können. (z.B. E-Mail-Adresse, Telefonnummer, usw.)

q7: Werden die erhobenen Daten entweder explizit oder zumindest in Form von Kategorien genannt?
Explizit: Wir erheben Ihren Namen, Wohnort, Straße, Postleitzahl. Kategorien: Wir erheben Adressdaten.
Der Anbieter ist verpflichtet den Nutzer über eine Datenerhebung zu unterrichten. Hierbei hat er die Daten so genau zu bezeichnen, dass es für den Nutzer möglich ist genau nachzuvollziehen, um welche Daten es sich handelt. Zu ungenau und folglich nicht ausreichend wäre also zum Beispiel: Wir erheben Daten zu Abrechnungszwecken.

q8: Wird genannt bei welcher Gelegenheit (Registrierung etc.) personenbezogene Daten erhoben und gespeichert werden?
Wir speichern Ihre Daten nach Beendigung der Registrierung.
Der Anbieter hat auf den Zeitpunkt der Datenerhebung hinzuweisen, damit dieser für den Nutzer nachvollziehbar ist.

q9: Wird genannt wie lange die erhobenen Daten gespeichert werden?
Wir löschen alle Ihre Daten, wenn Sie ihren Account löschen.
Denkbar ist auch eine konkrete Zeitangabe oder ein Hinweis auf die Sperrung der Daten anstelle ihrer Löschung.

q10: Wird genannt für welchen Zweck Daten erhoben werden?
Ihre Daten werden für Abrechnungszwecke, zur Kontaktaufnahme,... erhoben.
Der Anbieter ist verpflichtet, darauf hinzuweisen, zu welchem Zweck er Daten der Nutzer speichert.

q11: Findet die Unterrichtung der Datenerhebung vor dem Abschluss der Registrierung statt?
Der Anbieter ist verpflichtet, dem Nutzer die Informationen über die Datenerhebung und -verwendung spätestens zu Beginn der Dienstnutzung zukommen zu lassen.
Der Anbieter ist verpflichtet, den Nutzer vor der Nutzung über die Datenschutzbestimmungen zu unterrichten, damit sich dieser freiwillig für oder gegen diese entscheiden kann. Dafür reicht es aus, wenn bereits auf der ersten Seite des Anbieters ein Hinweis zur Datenschutzerklärung angebracht ist.

Cookies

q12: Setzt der Anbieter Cookies ein?
Cookies werden von CLEF automatisiert erkannt. Werden in obiger Live-Statistik Cookies angezeigt, setzt das Unternehmen auch Cookies ein.
Cookies sind kleine Dateien, die beim Besucher einer Webseite gespeichert werden. Zumeist beinhaltet das Cookie nur eine Nummer, die es dem Anbieter erlaubt, innerhalb einer Session (die Session endet mit dem Schließen des Browsers), oder aber auch über längere Zeit (persistente Cookies) Rückschlüsse zu ziehen, wer, wann, welche Seite besucht hat.

q13: Wird der Zweck des Einsatzes von Cookies genannt?
Wir erstellen Cookies zur Realisierung eines Online-Warenkorbs; zum Speichern von grafischen Einstellungen einer Webseite;
Der Anbieter ist verpflichtet, den Nutzer über eine Datenerhebung zu informieren. Hierbei ist auch auf den Zweck der Speicherung hinzuweisen.

q14: Wird der Zeitpunkt genannt, wann über Cookies Daten auf dem Benutzersystem gespeichert oder Benutzerdaten an den Anbieter gesendet werden?
Wenn der Benutzer die Webseite betritt, sich registriert, die erste Unterseite anwählt.
Der Anbieter ist verpflichtet, den Nutzer spätestens zu Beginn des Nutzungsvorgangs auch darüber zu informieren, wann automatisiert Daten erhoben und verwendet werden.

q15: Setzt der Anbieter Session Cookies ein?
(siehe Eintrag in der Live Statistik)
Session Cookies sind kleine Dateien, die auf dem Nutzer-Rechner gespeichert, aber nach Beenden einer Sitzung, d.h., nach dem Schließen des Browsers, automatisch gelöscht werden.

q16: Weist der Anbieter in der Datenschutzerklärung auf den Einsatz von Cookies hin?
Wir setzen Cookies ein, um Ihnen die Navigation zu vereinfachen.

q17: Setzt der Anbieter persistente Cookies ein?
(siehe Eintrag in der Live Statistik)
Persistente Cookies sind kleine Dateien, die auf dem Nutzer-Rechner gespeichert werden. Diese werden als persistent bezeichnet, wenn sie länger als für den Zeitraum einer Sitzung (Session) gespeichert werden.

q18: Weist der Anbieter in der Datenschutzerklärung auf den Einsatz von persistenten Cookies hin?
Wir setzen Cookies ein, die längerfristig auf Ihrem Computer gespeichert werden.
Persistente Cookies sind kleine Dateien, die auf dem Nutzer-Rechner gespeichert werden. Diese werden als persistent (permanent/langfristig gespeichert) bezeichnet, wenn sie länger als für den Zeitraum einer Sitzung (Session) gespeichert werden.

q19: Wird der konkrete Zeitraum genannt, für den Cookies auf dem Benutzersystem gespeichert werden?
Wir speichern Cookies 10 Jahre.

q20: Stimmt die vom Anbieter genannte Speicherdauer der Cookies mit der tatsächlichen Speicherdauer überein?
In der Datenschutzerklärung steht, dass Cookies 12 Monate gespeichert werden, tatsächlich werden sie aber 24 Monate speichert.
Die Informationen hierzu sind zu Beginn des blauen Bogenabschnitts zu finden (Live-Daten).

Webanalysedienste

q21: Setzt der Anbieter Webbugs, Analysedienste oder WebTracker ein?
Webbugs, Analysedienste und Tracker werden von CLEF automatisiert erkannt. Werden in obiger Live-Statistik Webbugs, Analysedienste oder Tracker angezeigt, setzt das Unternehmen diese auch ein.
WebBugs, Analysedienste und WebTracker ermöglichen es Anbietern, innerhalb ihrer Webseite, aber auch über unterschiedliche Webseiten hinweg, zu erkunden, von wo ein Benutzer kam, welche Seiten er besucht hat, und auf welche Seiten er weitergezogen ist.

q22: Setzt der Anbieter Google Analytics ein?
(siehe Eintrag in der Live Statistik)

q23: Wird in der Datenschutzerklärung auf den Einsatz von Webbugs, Analysediensten oder WebTrackern hingewiesen?
Wir setzen Google Analytics ein.

Pseudonymisierte Profile

q24: Weist der Anbieter auf die Erstellung von pseudonymisierten Nutzungsprofilen hin (explizit oder gibt er an, Browsertyp, IP Adresse, etc. zu speichern)?
Zu Statistikzwecken erstellen wir pseudonymisierte Nutzungsprofile.
Alternativ deutet die Erhebung des Betriebssystems, des Browsers etc. auf die Erstellung pseudonymisierter Nutzungsprofile hin.

q25: Wird auf den Zweck der Nutzungsprofile hingewiesen?
Webstatistik, Werbung
Der Anbieter ist bei einer Datenerhebung verpflichtet darauf hinzuweisen, zu welchem Zweck er Daten speichert.

q26: Ist der Zweck einer der folgenden: Werbung, Marktforschung oder die bedarfsgerechte Gestaltung der Telemedien?
Wir nutzen Nutzungsprofile, um Ihnen eine grafisch angepasste Gestaltung unserer Dienste zu ermöglichen.

q27: Wird der Benutzer darauf hingewiesen, dass er der Erstellung/Nutzung pseudonymisierter Profile widersprechen kann?
Der Erstellung und Speicherung von Nutzungsprofilen kann jederzeit mit Wirkung für die Zukunft widersprochen werden.
Der Nutzer hat jederzeit die Möglichkeit, der Datenverarbeitung zu widersprechen. Hierauf ist er vom Anbieter hinzuweisen.

q28: Ist es (gemäß der Datenschutzerklärung) möglich, auf elektronische Art (Klick auf Schaltfläche, Häkchen setzen, E-Mail, Kontaktformular) der Erstellung pseudonymisierter Profile zu widersprechen?
Über unten ausgewiesenes Kontaktformular können Sie dagegen Widerspruch einlegen.
Der Widerspruch muss ohne Medienbruch erklärt werden können. Handelt es sich um ein elektronisches Kommunikations- und Informationsmedium (Telemedium), muss auch der Widerspruch elektronisch erklärt werden können. Es ist nicht zumutbar, dass der Nutzer postalisch seinen Widerspruch erklärt.

Datenweitergabe

q29: Werden personenbezogene Daten (z.B. aus der Registrierung) an andere Unternehmen weitergegeben?
Wir übermitteln ihre Daten zu Abrechnungszwecken an die Abrechnungs AG. NICHT gemeint sind Daten aus Web Statistik Programmen wie Google Analytics.
Weitergabe bedeutet Datenübermittlung. Diese ist eine Datenverwendung, über die der Nutzer zu Beginn der Nutzung zu unterrichten ist.

q30: Werden die Empfänger-Unternehmen explizit genannt?
Wir geben Ihre Daten an die XY AG weiter.
Der Anbieter hat den Nutzer möglichst genau über die Empfänger der Daten zu informieren und den Zweck der Datenübermittlung anzugeben. Es reicht dabei NICHT aus, nur die Kategorien von Empfängern der Daten zu nennen (Bsp: Kreditinstitute,Transportunternehmen,...).

q31: Befindet sich das Unternehmen, an das die Daten weitergegeben werden, außerhalb der Europäischen Union (EU)?
China, Australien, Bolivien.
Mitgliedstaaten der Europäischen Union: Belgien, Italien, Rumänien, Bulgarien, Lettland, Schweden, Dänemark, Litauen, Slowakei, Deutschland, Luxemburg, Slowenien, Estland, Malta, Spanien, Finnland, Niederlande, Tschechien, Frankreich, Österreich, Ungarn, Griechenland, Polen, Vereinigtes Königreich, Irland, Portugal, Republik Zypern

q32: Handelt es sich bei dem Land um: Argentinien, Guernsey, Isle of Man, Kanada oder die Schweiz?
Argentinien.
Mit diesen Staaten/Territorien bestehen spezielle Abkommen.

q33: Ist das Unternehmen, an das die Daten weitergegeben werden, Mitglied des Safe Harbor Abkommens?
Microsoft, HP, Google, General Motors, Amazon.
Informationen, ob das Unternehmen Mitglied des Safe Harbor-Abkommens ist, finden sie in der Regel in der Datenschutzbestimmung, in den AGB.

q34: Verwenden diese Unternehmen laut Datenschutzerklärung die Daten für einen anderen Zweck als die Geschäftsabwicklung oder die Abrechnung?
Für Werbezwecke, zur Kontaktaufnahme.
Nur für die Geschäfts-/ Vertragsabwicklung und für die Abrechnung dürfen Nutzungsdaten ohne Einwilligung an eine andere verantwortliche Stelle übermittelt werden. In jedem anderen Fall ist eine Einwilligung erforderlich.

Aktionen, die Einwilligung erfordern

q35: Wird bei der Registrierung mehr als eine Möglichkeit zur direkten Kontaktaufnahme abgefragt?
Pflichtfelder: E-Mail-Adresse und Handynummer.
Die Erhebung EINES Kontaktdatums bedarf im Regelfall keiner Einwilligung. Sie ist gesetzlich legitimiert. Werden weitere Kontaktdaten erhoben (beispielsweise neben der E-Mail-Adresse auch eine Telefonnummer), so bedarf es hierfür der Einholung einer Einwilligungserklärung, da regelmäßig das zusätzliche Datum für die Kontaktaufnahme nicht erforderlich ist.

q36: Wird in der Datenschutzerklärung darauf hingewiesen, dass personalisierte Benutzerprofile erstellt werden?
Wir erstellen personalisierte Benutzerprofile, um Ihnen auch andere Produkte anzeigen zu können, an denen Sie Interesse haben könnten.
Alternativ werden diese Profile auch individualisiert, 'auf Sie zugeschnitten' etc. genannt. Schreibt der Anbieter, dass KEINE personalisierten Profile erstellt werden, beantworten Sie die Frage mit NEIN.

q37: Wird eine Einwilligung vom Benutzer eingeholt?
Ein Häkchen im Registrierungsformular und ein Text ähnlich 'Ich willige ein, dass...' in der Datenschutzerklärung.

q38: Muss man in den AGB in die Datenschutzpraktik einwilligen UND ist der Teil der Einwilligung in den AGB grafisch hervorgehoben?
Antworten Sie nur dann mit NEIN, wenn eine Einwilligung innerhalb der AGB gegeben werden muss UND gleichzeitig KEINE grafische Hervorhebung existiert. Ansonsten antworten Sie bitte mit JA!
Der Anbieter ist verpflichtet, die Einwilligungserklärung, wenn sie zusammen mit anderen Erklärungen erteilt wird, besonders hervorzuheben. Dieser Pflicht kann er mit Hilfe optischer Mittel nachkommen, z.B. Fettdruck, farbliche Markierung,...

q39: Ist die Einwilligungserklärung jederzeit abrufbar?
Als Teil der Datenschutzerklärung/AGB oder als herunterladbare Datei?
Der Anbieter ist verpflichtet, sicherzustellen, dass der Nutzer jederzeit den Inhalt seiner Einwilligung abrufen kann.

q40: Wird in der Einwilligungserklärung darauf hingewiesen, dass die Einwilligung durch den Benutzer jederzeit mit Wirkung für die Zukunft widerrufbar ist?
Die Einwilligung ist jederzeit mit Wirkung für die Zukunft widerrufbar.
Der Nutzer hat das Recht, seine Einwilligung jederzeit zu widerrufen. Hierauf hat der Anbieter explizit hinzuweisen.

q41: Muss der Benutzer in Forderungen einwilligen, die im Konflikt mit anderen Datenschutzartikeln stehen?
Ich willige ein, dass die Test AG mir keine Auskünfte über die bei ihr gespeicherten Daten geben muss.

D.4. Resultierende Verstöße der Taxonomie

Dieser Abschnitt beinhaltet alle Verstöße, die wir für CLEF modelliert haben. Die Verstöße sind wiederum thematisch aufgeteilt. Ein Verstoß besteht aus einer eindeutigen Identifikationsnummer (*VerstoßID*) und einem Verstoßnamen. Dazu kommen das Antwortmuster und der Gesetzesparagraph, aus dem sich der Verstoß ableitet.

VerstoßID: Verstoßname
Antwortmuster
Gesetzesparagraph

Datenschutzerklärung

v1: Es existiert keine Datenschutzerklärung.
$\neg q1$
§ 13 TMG

v2: Die Datenschutzerklärung ist nicht jederzeit abrufbar.
$q1 \wedge \neg q2$
§ 13 Abs. 1 S. 3 TMG

v3: In der Datenschutzerklärung ist keine Kontakt-Adresse des Anbieters genannt.
$q1 \wedge \neg q3$
§ 5 Abs. 1 Nr. 1 TMG

v4: In der Datenschutzerklärung ist keine Kontakt-E-Mail-Adresse des Anbieters genannt.
$q1 \wedge \neg q4$
§ 5 Abs. 1 Nr. 2 TMG

v5: In der Datenschutzerklärung wird nicht das Handelsregister des Anbieters genannt.
$q1 \wedge \neg q5$
§ 5 Abs. 1 Nr. 4 TMG

Datenerhebung bei Registrierung

v6: Kein Hinweis, dass personenbezogene Daten vom Benutzer erhoben werden.
$\neg q6$
§ 13 Abs. 1 S. 1 TMG

v7: Kein Hinweis, welche Daten erhoben werden.
$q6 \wedge \neg q7$
§ 13 Abs. 1 S. 1 TMG

v8: Kein Hinweis, wann Daten erhoben werden.
$q6 \wedge \neg q8$
§ 13 Abs. 1 S. 1 TMG

v9: Kein Hinweis, wie lange Daten gespeichert werden.
$q6 \wedge \neg q9$
§ 13 Abs. 1 S. 1 TMG

v10: Kein Hinweis, weshalb Daten erhoben werden.
$q6 \wedge \neg q10$
§ 13 Abs. 1 S. 1 TMG

v11: Unterrichtung über Datenerhebung findet nicht zu Beginn des Nutzungsvorgangs statt.
$q6 \wedge q7 \wedge q8 \wedge q9 \wedge q10 \wedge \neg q11$
§ 13 Abs. 1 S. 1 TMG

Cookies

v12: Kein Hinweis auf den Einsatz von Cookies.
$q12 \wedge \neg q16$
§ 13 Abs. 1 TMG

v13: Kein Hinweis, wann Cookies gespeichert werden.
$q12 \wedge \neg q14$
§ 13 Abs. 1 TMG

v14: Kein Hinweis, für welchen Zweck Cookies eingesetzt werden.
$q12 \wedge \neg q13$
§ 13 Abs. 1 TMG

v15: Kein Hinweis, dass persistente Cookies eingesetzt werden.
$q12 \wedge q17 \wedge \neg q18$
§ 13 Abs. 1 TMG

v16: Kein Hinweis, wie lange Cookies gespeichert werden.
$q12 \wedge \neg q19$
§ 13 Abs. 1 TMG

v17: Genannte Speicherzeit der Cookies ist falsch.
$q12 \wedge q19 \wedge \neg q20$
§ 13 Abs. 1 TMG

Webanalysedienste

v18: Einsatz von Google Analytics.
$q22$
Beschluss Düsseldorfer Kreis November 2009

v19: Kein Hinweis auf Einsatz von Webbugs.
$q21 \wedge \neg q23$
§ 13 Abs. 1 TMG

Pseudonymisierte Profile

v20: Zweck der pseudonomisierten Verfahren nicht genannt.
$q24 \wedge \neg q25$
§§ 13 Abs. 1 S. 1, 15 Abs. 3 S. 1 TMG

v21: Zweck der pseudonomisierten Verfahren nicht zulässig.
$q24 \wedge q25 \wedge \neg q26 \wedge$ *Keine Einwilligung*
§§ 12 Abs. 1, 15 Abs. 3 S. 1 TMG

v22: Kein Hinwies auf Widerspruchsrecht für pseudonymisierte Profile.
$q24 \wedge \neg q27$
§§ 15 Abs. 3 S. 2, 13 Abs. 1 TMG

v23: Widerspruch der Erstellung pseudonymisierter Profile nicht ohne Medienbruch möglich.
$q24 \wedge q27 \wedge \neg q28$
§§ 15 Abs. 3 S.1, 13 Abs. 1 TMG

Datenweitergabe

v24: Empfänger der weitergegebenen Daten nicht genannt.
$q29 \wedge \neg q30$
§ 13 Abs. 1 TMG
v25: Keine Einwilligung in Weitergabe der Daten.
$q29 \wedge q31 \wedge \neg q32 \wedge \neg q33 \wedge$ *Keine Einwilligung*
§ 12 Abs. 1 TMG

Aktionen, die Einwilligung erfordern

v26: Keine Einholung der Einwilligung trotz einwilligungsbedürftigem Datenumgang.
$q35 \wedge q37$
§§ 12 Abs. 1, 3 TMG
v27: Keine Einwilligung in Erstellung personalisierter Profile.
$q36 \wedge q37$
§§ 12 Abs. 1 TMG
v28: Einwilligungstext nicht deutlich hervorgehoben.
$(q35 \vee q36) \wedge q37 \wedge \neg q38$
§§ 12 Abs. 1 TMG, 4a Abs. 1 S. 4 BDSG
29: Einwilligungstext nicht jederzeit abrufbar.
$(q35 \vee q36) \wedge q37 \wedge \neg q39$
§ 13 Abs. 1 S. 3 TMG
v30: Kein Hinweis auf Widerrufsrecht für die Einwilligung.
$(q35 \vee q36) \wedge q37 \wedge \neg q40$
§§ 13 Abs. 3, Abs. 2 Nr. 4 TMG
v31: Einwilligung in unzulässige Forderungen.
$q41$
§ 12 Abs. 1 TMG

E. Anonymisierung von Suchhistorien

Dieser Anhang liefert weitere Details zu der Anonymisierung von Suchhistorien (Kapitel 6). Diese beziehen sich auf das Laufzeitverhalten unserer Heuristik (Abschnitt E.1) und den resultierenden Nutzen eines anonymisierten Suchprotokolls (Abschnitt E.2).

E.1. Iterationen pro Zielfunktion und k

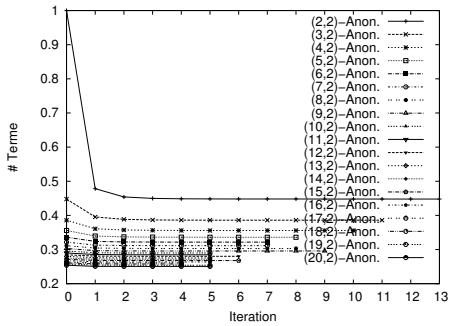


Abbildung E.1.: m=2: Protokollgröße / Iteration

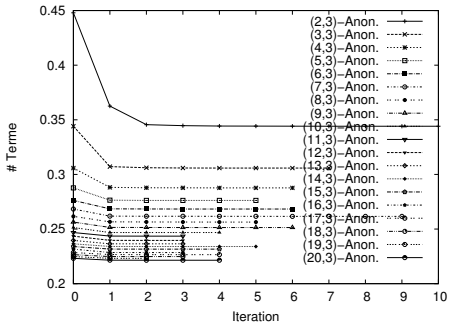


Abbildung E.2.: m=3: Protokollgröße / Iteration

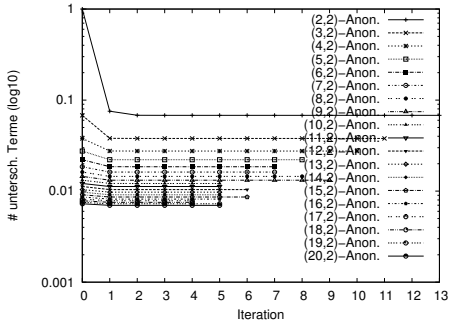


Abbildung E.3.: m=2: Untersch. Terme / Iteration

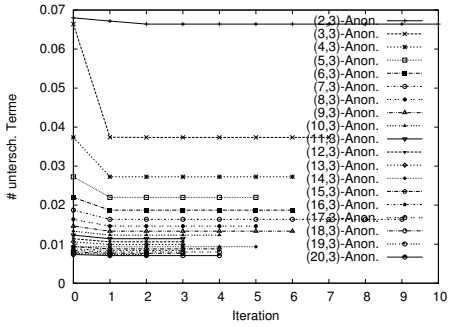


Abbildung E.4.: m=3: Untersch. Terme / Iteration

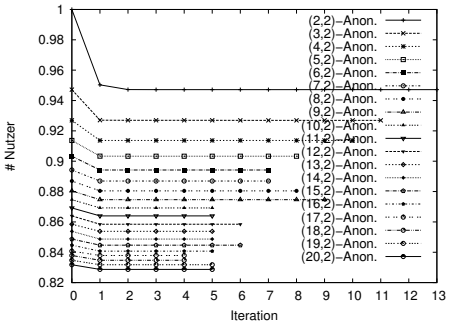


Abbildung E.5.: m=2: Nutzer/ Iteration

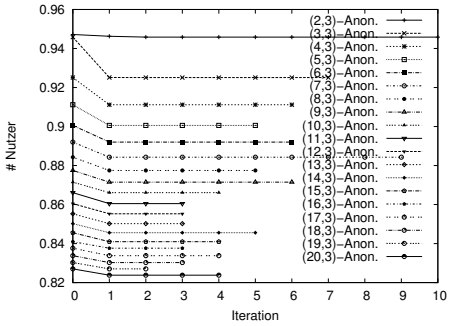


Abbildung E.6.: m=3: Nutzer/ Iteration

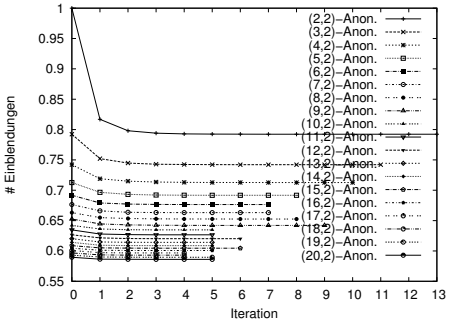


Abbildung E.7.: m=2: Einblendungen / Iteration

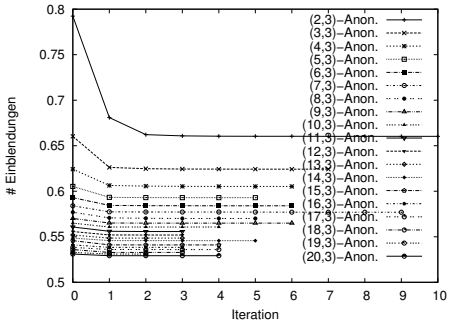


Abbildung E.8.: m=3: Einblendungen / Iteration

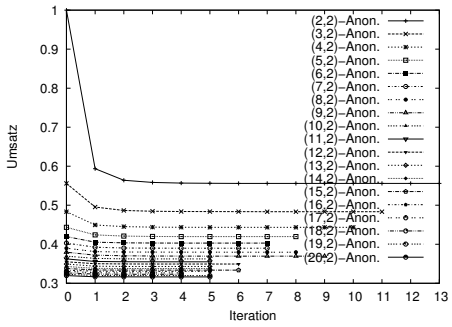


Abbildung E.9.: m=2: Umsatz/ Iteration

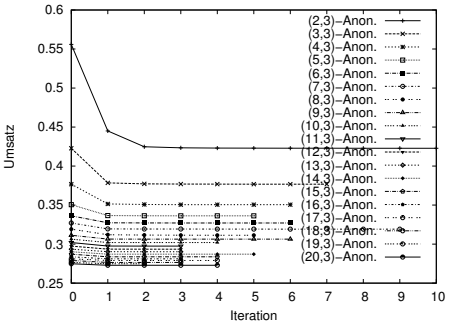


Abbildung E.10.: m=3: Umsatz / Iteration

E.2. Nutzen abhängig von der Zielfunktion

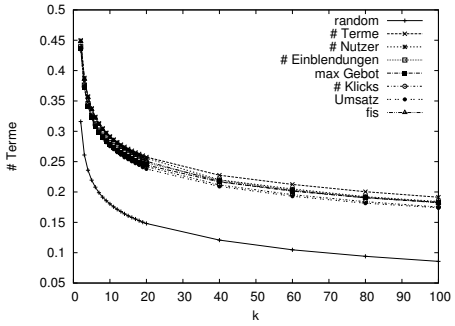


Abbildung E.11.: m=2: Protokollgröße / Zielfunktion

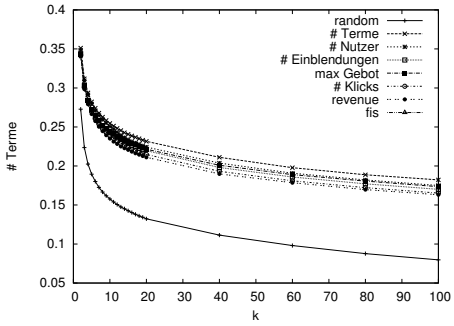


Abbildung E.12.: m=3: Protokollgröße / Zielfunktion

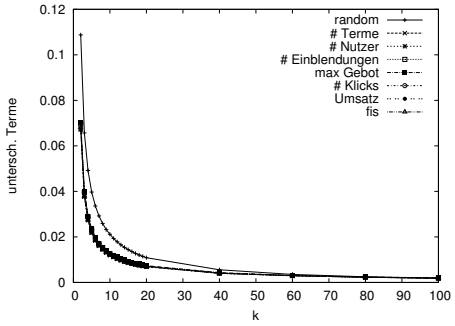


Abbildung E.13.: m=2: Untersch. Terme / Zielfunktion

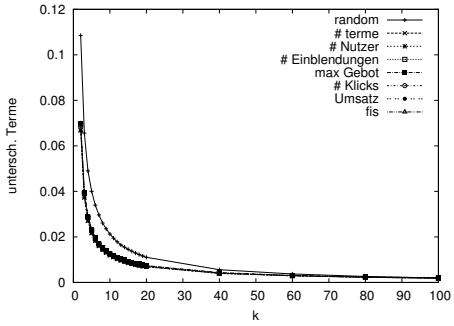


Abbildung E.14.: m=3: Untersch. Terme / Zielfunktion

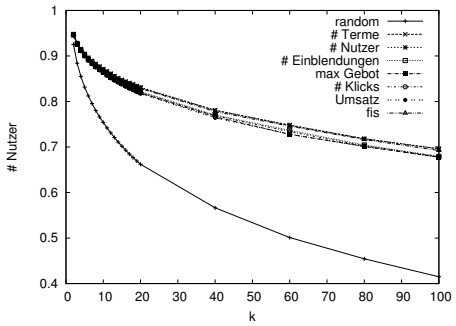


Abbildung E.15.: m=2: Nutzer / Zielfunktion

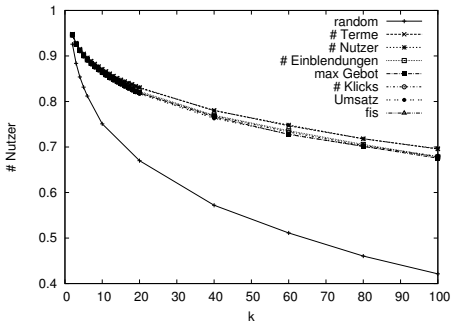


Abbildung E.16.: m=3: Nutzer / Zielfunktion

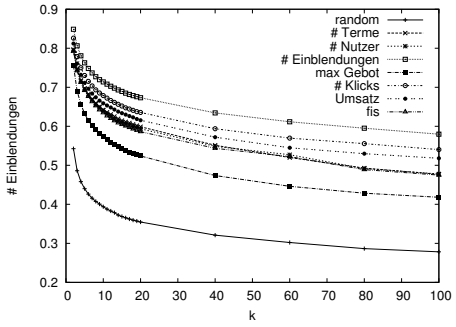


Abbildung E.17.: m=2: Einblendungen / Zielfunktion

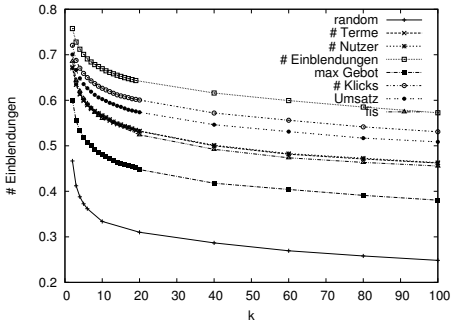


Abbildung E.18.: m=3: Einblendungen / Zielfunktion

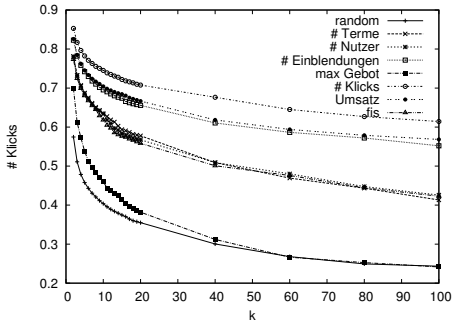


Abbildung E.19.: m=2: Klicks / Zielfunktion

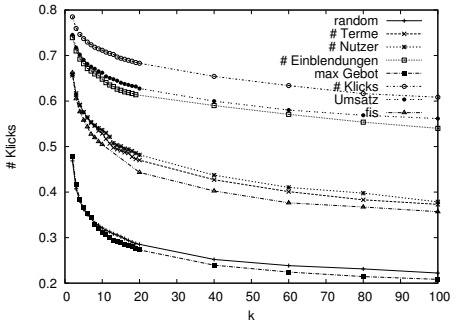


Abbildung E.20.: m=3: Klicks / Zielfunktion

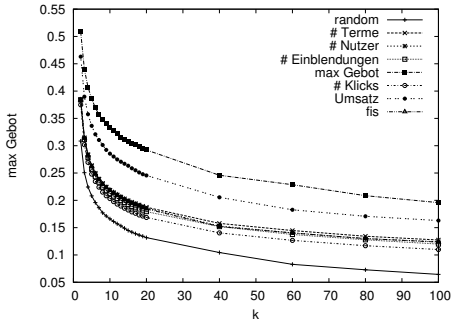


Abbildung E.21.: m=2: Gebot / Zielfunktion

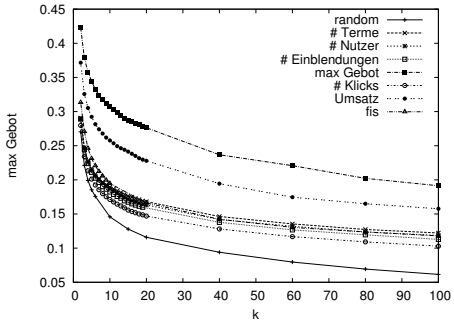


Abbildung E.22.: m=3: Gebot / Zielfunktion

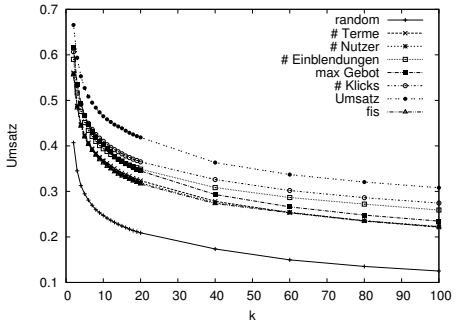


Abbildung E.23.: m=2: Umsatz / Zielfunktion

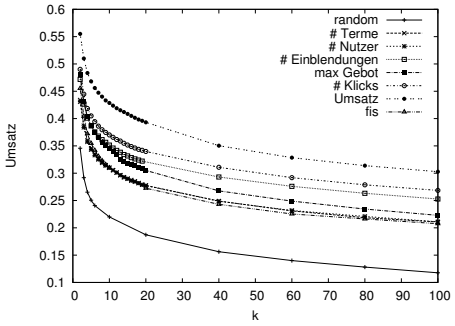


Abbildung E.24.: m=3: Umsatz / Zielfunktion

Literaturverzeichnis

F.1. Quellen

[Ack00] ACKERMAN, Mark S.: The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. In: *Human-Computer Interaction* 15 (2000), Nr. 2, S. 179–203. – ISSN 0737–0024

[ACR99] ACKERMAN, Mark S. ; CRANOR, Lorrie F. ; REAGLE, Joseph: Privacy in E-Commerce: Examining User Scenarios and Privacy Preferences. In: *EC '99: Proceedings of the 1st Conference on Electronic Commerce*, ACM, 1999. – ISBN 1–58113–176–3, S. 1–8

[Ada07] ADAR, Eytan: User 4xxxxx9: Anonymizing Query Logs. In: *Workshop on Query Log Analysis at World Wide Web Conference*, ACM, 2007

[AG05] ACQUISTI, Alessandro ; GROSSKLAGS, Jens: Privacy and Rationality in Individual Decision Making. In: *IEEE Security and Privacy* (2005)

[agi10] A Call for Agility: The Next-Generation Privacy Professional. In: *IAAP: International Association of Privacy Professionals*, 2010

[Aig10] AIGNER, Ilse: Offener Brief an Zuckerberg (Facebook), "Privates muss privat bleiben". (2010)

[AKSX03] AGRAWAL, Rakesh ; KIERNAN, Jerry ; SRIKANT, Ramakrishnan ; XU, Yirong: An XPath-based Preference Language for P3P. In: *WWW '03: Proceedings of the 12th International Conference on World Wide Web*, ACM, 2003. – ISBN 1–58113–680–3, S. 629–639

[AN07] AMES, Morgan ; NAAMAN, Mor: Why we tag: Motivations for Annotation in Mobile and Online Media. In: *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2007. – ISBN 978–1–59593–593–9, S. 971–980

[And09] ANDERSON, Chris: *Free, the Future of a Radical Price*. Campus Verlag, 2009. – ISBN 978–1401322908

[Arb08] ARBEITSGEMEINSCHAFT ONLINE-FORSCHUNG E.V: *ÄGOF Internet Facts 2008-IV"*. http://www.agof.de, 2008

[Bab07] BABBIE, Earl R.: *The Practice of Social Research*. 10. Academic Internet Publ., 2007

[BBR08] BUCHMANN, Erik ; BOEHM, Klemens ; RAABE, Oliver: Privacy2.0: Towards Collaborative Data-Privacy Protection. In: *Proceeding of IFIPTM08*, 2008

[BD03] BARKHUUS, Louise ; DEY, Anind: Location-Based Services for Mobile Telephony: A Study of Users' Privacy Concerns. In: *INTERACT '03: Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction*, 2003, S. 709–712

[BGD+06] BIRD, Christian ; GOURLEY, Alex ; DEVANBU, Prem ; GERTZ, Michael ; SWAMINATHAN, Anand: Mining Email Social Networks. In: *MSR '06: Proceedings of the 2006 International Workshop on Mining Software Repositories*, ACM, 2006. – ISBN 1–59593–397–2, S. 137–143

[BKK06] BRODIE, Carolyn A. ; KARAT, Clare-Marie ; KARAT, John: An Empirical Study of Natural Language Parsing of Privacy Policy Rules using the SPARCLE Policy Workbench. In: *SOUPS '06: Proceedings of the second Symposium on Usable Privacy and Security*, ACM, 2006, S. 8–19

[BM02] BAYE, Michael R. ; MORGAN, John: Information gatekeepers and Price Discrimination on the Internet. In: *Economics Letters* 76 (2002), Nr. 1, S. 47 – 51. – ISSN 0165–1765

[BM04] BARGH, J. A. ; MCKENNA, K. Y.: The internet and social life. In: *Annual Review of Psychology* 55 (2004), S. 573–590

[Bor05a] BORGELT, Christian: An Implementation of the FP-growth Algorithm. In: *OSDM '05: Proceedings of the 1st International Workshop on Open Source Data Mining*, ACM, 2005. – ISBN 1–59593–210–0, S. 1–5

[Bor05b] BORTZ, Jürgen: *Statistik für Human- und Sozialwissenschaftler*. Springer, 2005. – ISBN 978–3–642–12769–4

[BRDM07] BEATTY, Patricia ; REAY, Ian ; DICK, Scott ; MILLER, James: P3P Adoption on E-Commerce Web sites: A Survey and Analysis. In: *IEEE Internet Computing* 11 (2007), Nr. 2, S. 65–71. – ISSN 1089–7801

[BTZ06] BARBARO, Michael ; TOM ZELLER, Jr.: A Face Is Exposed for AOL Searcher NO. 4417749. In: *NYT* (2006). `http://www.nytimes.com/2006/08/09/technology/09aol.html`

[CGA06] CRANOR, Lorrie F. ; GUDURU, Praveen ; ARJULA, Manjula: User Interfaces for Privacy Agents. In: *Computer-Human Interaction* 13 (2006), Nr. 2, S. 135–178. – ISSN 1073–0516

[CKMD06] CVRCEK, Dan ; KUMPOST, Marek ; MATYAS, Vashek ; DANEZIS, George: A study on the value of location privacy. In: *WPES '06: Proceedings of the 5th Workshop on Privacy in Electronic Society*, ACM, 2006. – ISBN 1–59593–556–8, S. 109–118

[Cla71] CLARKE, Edward H.: Multipart Pricing of Public Goods. In: *Public Choice* 11 (1971), September, Nr. 1. http://jmvidal.cse.sc.edu/library/clarke71a.pdf ISSN 1573–7101

[CLM] CRANOR, Lorrie ; LANGHEINRICH, Marc ; MARCHIONI, Massimo: A P3P Preference Exchange Language 1.0 (APPEL1.0). http://www.w3.org/TR/P3P-preferences/

[Coh60] COHEN, J.: A Coefficient of Agreement for Nominal Scales. In: *Educational and Psychological Measurement* 20 (1960)

[CSM+05] CONSOLVO, Sunny ; SMITH, Ian E. ; MATTHEWS, Tara ; LAMARCA, Anthony ; TABERT, Jason ; POWLEDGE, Pauline: Location Disclosure to Social Relations: Why, When, & What people want to share. In: *CHI '05: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2005. – ISBN 1–58113–998–5, S. 81–90

[Dat08] DATA PROTECTION WORKING PARTY: *Article 29 Opinion 1/2008 on data protection issues related to search engines*. 4 April 2008

[Deu10] DEUTSCHER BUNDESTAG, WISSENSCHAFTLICHE DIENSTE: *Aktueller Begriff Mobiles Internet*. http://www.bundestag.de/dokumente/analysen/2010/mobiles_internet.pdf. Version: 2010

[ECC06] EGELMAN, Serge ; CRANOR, Lorrie F. ; CHOWDHURY, Abdur: An Analysis of P3P-enabled Web Sites among top-20 Search Results. In: *ICEC '06: Proceedings of the 8th International Conference on Electronic Commerce*, ACM, 2006. – ISBN 1–59593–392–1, S. 197–207

[EF08] EIMEREN, Birgit von ; FREES, Beate: Medienperspektiven, ARD/ZDF Onlinestudie, 2008

[EGH08] EBERSBACH, A. ; GLASER, M. ; HEIGL, R.: *Social Web*. Utb, 2008. – ISBN 978–3825230654

[Eng05] ENGISCH, Karl: *Einführung in das juristische Denken*. 10. Aufl. (hrsg. Thomas Würtenberger). Kohlhammer, 2005 (Urban-Taschenbücher ; 20). – ISBN 3–17–018695–7

[EOS07] EDELMAN, B. ; OSTROVSKY, M. ; SCHWARZ, M.: Internet Advertising and the Generalized Second-Price Auction: Selling billions of dollars worth of keywords. In: *American Economic Review* 97 (2007), Nr. 1, S. 242–259

[Eur81] EUROPARAT: *Übereinkommen zum Schutz des Menschen bei der automatischen Verarbeitung personenbezogener Daten*. Straßburg, 1981

[Fac10] FACEBOOK PRESS ROOM: *Statistics*. http://www.facebook.com/press/info.php?statistics, Version: 01 2010

[FE08] FELT, Adrienne ; EVANS, David: Privacy Protection for Social Networking Platforms. In: *Web 2.0 Security and Privacy Workshop*, 2008

[Fel98] FELLBAUM, Christiane (Hrsg.): *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998. – ISBN 978–0262061971

[Fel07] FELT, Adrienne: *Defacing Facebook: A Security Case Study*. http://www.cs.virginia.edu/felt/fbook/facebook-xss.pdf, 2007

[GA05] GROSS, Ralph ; ACQUISTI, Alessandro: Information Revelation and Privacy in Online Social Networks. In: *WPES '05: Proceedings of the 2005 Workshop on Privacy in the Electronic Society*, ACM, 2005. – ISBN 1–59593–228–3, S. 71–80

[GG03] GRUTESER, Marco ; GRUNWALD, Dirk: Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In: *MobiSys '03: Proceedings of the 1st International Conference on Mobile Systems, Applications and Services*, ACM, 2003, S. 31–42

[GL04] GEDIK, Bugra ; LIU, Ling: A Customizable k-Anonymity Model for Protecting Location Privacy / Georgia Institute of Technology. 2004. – Forschungsbericht

[Gla01] GLANCE, Natalie S.: Community Search Assistant. In: *IUI '01: Proceedings of the 6th International Conference on Intelligent User Interfaces*, ACM, 2001. – ISBN 1–58113–325–1, S. 91–96

[Gol06] GOLLE, Philippe: Revisiting the Uniqueness of Simple Demographics in the US Population. In: *WPES '06: Proceedings of the 5th Workshop on Privacy in Electronic Society*, ACM, 2006. – ISBN 1–59593–556–8, S. 77–80

[Hau01] HAUBEN, Ronda: Cultural Clash – The Google Purchase of the 1995-2001 Usenet Archive And the Online Community. In: *Heise – Telepolis* (2001)

[HBB10] HEIDINGER, Clemens ; BUCHMANN, Erik ; BÖHM, Klemens: Collaborative Data Privacy for the Web. In: *EDBT '10: Proceedings of the 2010 EDBT Workshops*, ACM, 2010. – ISBN 978–1–60558–990–9, S. 1–10

[HJS] HART, M. ; JOHNSON, R. ; STENT, A.: More Content – Less Control: Access Control in the Web 2.0. In: *W2SP '07: Proceedings of the Web 2.0 Security and Privacy Workshop*, 43–48

[HN09] HE, Yeye ; NAUGHTON, Jeff: Anonymization of Set-Valued Data via Top-Down, Local Generalization. In: *VLDB '09: Proceedings of the 13th International Conference on Very Large Data Bases*, VLDB Endowment, 2009

[Hog02] HOGBEN, Giles: A technical Analysis of Problems with P3P v1.0 and Possible Solutions. Joint Research Centre, Dulles, Virginia, 2002. – Forschungsbericht

[Hoo05] HOOFNAGLE, Chris J.: Privacy Self Regulation: A decade of Disappointment. In: *Electronic Privacy Information Center* (2005)

[Hoo07] HOOFNAGLE, Chris J.: Identity Theft: Making the Known Unknowns Known. In: *Harvard Journal of Law and Technology* 21 (2007)

[HR06] HAFNER, Katie ; RICHTEL, Matt: Google Resists U.S. Subpoena of Search Data. In: *The New York Times* (2006), Januar 20. http://www.nytimes.com/2006/01/20/technology/20google.html

[HZNR09] HOFFMAN, Kevin ; ZAGE, David ; NITA-ROTARU, Cristina: A Survey of Attack and Defense Techniques for Reputation Systems. In: *ACM Computing Surveys* 42 (2009), Nr. 1

[JIB07] JØSANG, Audun ; ISMAIL, Roslan ; BOYD, Colin: A Survey of Trust and Reputation Systems for Online Service Provision. In: *Decision Support Systems* (2007)

[JKPT07] JONES, Rosie ; KUMAR, Ravi ; PANG, Bo ; TOMKINS, Andrew: Ï know what you did last summer": Query Logs and User Privacy. In: *CIKM '07: Proceedings of the 16th ACM Conference on Information and Knowledge Management*, ACM, 2007. – ISBN 978–1–59593–803–9, S. 909–914

[JL.71] JL., Fleiss: Measuring Nominal Scale Agreement among many Raters. In: *Psychol Bull*, 1971

[JSP05] JANSEN, Bernard J. ; SPINK, Amanda ; PEDERSEN, Jan: A temporal Comparison of AltaVista Web Searching: Research Articles. In: *Journal of the American Society for Information Science and Technology* 56 (2005), Nr. 6, S. 559–570. – ISSN 1532–2882

[JZ05] JUTLA, Dawn ; ZHANG, Yanjun: Maturing e-Privacy with P3P and Context Agents. In: *EEE '05: Proceedings of the International Conference on e-Technology, e-Commerce and e-Service*, IEEE, 2005, S. 536–541

[Kan06] KANTOR, Andrew: AOL search data release reveals a great deal. In: *USA Today* (2006), August 17. http://www.usatoday.com/tech/columnist/andrewkantor/2006-08-17-aol-data_x.htm

[KB10] KÜHLING, Jürgen ; BOHNEN, Simon: Zur Zukunft des Datenschutzrechts – Nach der Reform ist vor der Reform. In: *Juristen Zeitung (JZ)* (2010)

[KF07] KRIEWEL, Sascha ; FUHR, Norbert: Adaptive Search Suggestions for Digital Libraries. In: *ICADL'07: Proceedings of the 10th International Conference on Asian Digital Libraries*, Springer-Verlag, 2007. – ISBN 3–540–77093–3, 978–3–540–77093–0, S. 220–229

[KKMN09] KOROLOVA, Aleksandra ; KENTHAPADI, Krishnaram ; MISHRA, Nina ; NTOULAS, Alexandros: Releasing Search Qqueries and Clicks Privately. In: *WWW '09: Proceedings of the 18th International Conference on World Wide Web*, ACM, 2009. – ISBN 978–1–60558–487–4, S. 171–180

[Kno03] KNORR, E.: The Year of Web Services: The Stars Will Align in 2004 to Make Web Services a Significant Success Story. In: *CIO* 15 (2003), 12, S. 90

[KNPT07] KUMAR, Ravi ; NOVAK, Jasmine ; PANG, Bo ; TOMKINS, Andrew: On Anonymizing Query Logs via Token-based Hashing. In: *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, ACM, 2007. – ISBN 978–1–59593–654–7, 629–638

[KSS08] KÜHLING, Jürgen ; SEIDEL, Christian ; SIVRIDIS, Anastasios: *Datenschutzrecht*. Frankfurt am Main : UTB, 2008 (UTB ; 3109). – ISBN 978–3–8252–3109–5

[KYS05] KIDO, Hidetoshi ; YANAGISAWA, Yutaka ; SATOH, Tetsuji: Protection of Location Privacy using Dummies for Location-based Services. In: *ICDEW '05: Proceedings of the 21st International Conference on Data Engineering Workshops*, IEEE, 2005. – ISBN 0–7695–2657–8, S. 1248

[LB08] LUCAS, Matthew M. ; BORISOV, Nikita: FlyByNight: Mitigating the Privacy Risks of Social Networking. In: *WPES '08: Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society*, ACM, 2008. – ISBN 978–1–60558–289–4, S. 1–8

[LBR02] LARGE, Andrew ; BEHESHTI, Jamshid ; RAHMAN, Tarjin: Gender Differences in Collaborative Web Searching Behavior: an Elementary School Study. In: *Information Processing and Management* 38 (2002), Nr. 3, S. 427–443. – ISSN 0306–4573

[LH08] LESKOVEC, Jure ; HORVITZ, Eric: Planetary-scale Views on a large Instant-Messaging Network. In: *WWW '08: Proceeding of the 17th International Conference on World Wide Web*, ACM, 2008. – ISBN 978–1–60558–085–2, S. 915–924

[LHU09] LANDESANSTANT FÜR MEDIEN ; HANS-BREDOW-INSTITUT ; UNIVERSITÄT SALZBURG: Heranwachsen mit dem Social Web – Zur Rolle von Web 2.0 Angeboten im Alltag von Jugendlichen und jungen Erwachsenen, Landesanstalt für Medien Nordrhein-Westfalen (LfM), 2009

[LK77] LANDIS, J. R. ; KOCH, G. G.: The Measurement of Observer Agreement for Categorical Data. In: *Biometrics* 33 (1977), March, Nr. 1, S. 159–174. – ISSN 0006–341X

[LLV] LI, Ninghui ; LI, Tiancheng ; VENKATASUBRAMANIAN, Suresh: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, IEEE

[LMD03] LEDERER, Scott ; MANKOFF, Jennifer ; DEY, Anind K.: Who wants to know What When? Privacy Preference Determinants in Ubiquitous Computing. In: *CHI '03: CHI '03 extended abstracts on Human Factors in Computing Systems*, ACM, 2003. – ISBN 1–58113–637–4, S. 724–725

[LSY03] LINDEN, Greg ; SMITH, Brent ; YORK, Jeremy: Amazon.com Recommendations: Item-to-Item Collaborative Filtering. In: *IEEE Internet Computing* 7 (2003), S. 76–80. – ISSN 1089–7801

[Mar02] MARCHIORI, M.: *The Platform for Privacy Preferences (P3P 1.0) Specification*. W3C Proposed Recommendation, 2002

[Mat04] MATHES, A.: Folksonomies – Cooperative Classification and Communication through Shared Metadata. In: *Computer Mediated Communication* (2004)

[MC08a] MCDONALD, Aleecia M. ; CRANOR, Lorrie F.: The Cost of Reading Privacy Policies. In: *A Journal of Law and Policy for the Information Society* (2008)

[MC08b] MURUGESAN, Mummoorthy ; CLIFTON, Chris: Plausibly Deniable Search. In: *Secure Knowledge Management*, 2008

[MCA06] MOKBEL, Mohamed F. ; CHOW, Chi-Yin ; AREF, Walid G.: The new Casper: Query Processing for Location Services without Compromising Privacy. In: *VLDB '06: Proceedings of the 32nd International Conference on Very Large Data Bases*, VLDB Endowment, 2006, S. 763–774

[Med06a] MEDIENPÄDAGOGISCHER FORSCHUNGSVERBAND SÜDWEST (MPFS): *JIM-Studie - Jugend, Information, Multimedia* . `http://www.mpfs.de/fileadmin/JIM--pdf06/JIM--Studie_2006.pdf`, 2006

[Med06b] MEDIENPÄDAGOGISCHER FORSCHUNGSVERBAND SÜDWEST (MPFS): *KIM-Studie - Kinder + Medien, Computer + Internet*. http://www.mpfs.de/fileadmin/Studien/KIM05.pdf, Version: 2006

[MGKV06] MACHANAVAJJHALA, Ashwin ; GEHRKE, Johannes ; KIFER, Daniel ; VENKITASUBRAMANIAM, Muthuramakrishnan: l-Diversity: Privacy Beyond k-Anonymity. In: *ICDE '06: 22nd IEEE International Conference on Data Engineering*, IEEE, 2006

[MH07] MORRIS, Meredith R. ; HORVITZ, Eric: SearchTogether: An Interface for Collaborative Web Search. In: *UIST '07: Proceedings of the 20th annual ACM Symposium on User Interface Software and Technology*, ACM, 2007. – ISBN 978–1–59593–679–2, S. 3–12

[Mil95] MILLER, George A.: WordNet: a Lexical Database for English. In: *Commun. ACM* 38 (1995), Nr. 11, S. 39–41. – ISSN 0001–0782

[MNBD06] MARLOW, Cameron ; NAAMAN, Mor ; BOYD, Danah ; DAVIS, Marc: HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: *HYPERTEXT '06: Proceedings of the 17th Conference on Hypertext and Hypermedia*, ACM, 2006. – ISBN 1–59593–417–0, S. 31–40

[Mor07] MORRIS, Meredith R.: Collaborating Alone and Together: Investigating Persistent and Multi-User Web Search Activities / Microsoft Research. 2007. – Forschungsbericht

[MW04] MEYERSON, Adam ; WILLIAMS, Ryan: On the Complexity of optimal K-Anonymity. In: *PODS '04: Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ACM, 2004, S. 223–228

[Net06] NETCRAFT: *November 2006 Web Server Survey*. http://news.netcraft.com/archives/2006/11/01/november_2006_web_server_survey.html, Version: 2006

[NS09] NARAYANAN, Arvind ; SHMATIKOV, Vitaly: De-anonymizing Social Networks. In: *IEEE Symposium on Security and Privacy* (2009), S. 173–187. – ISSN 1081–6011

[NS10] NARAYANAN, Arvind ; SHMATIKOV, Vitaly: Myths and Fallacies of "Personally Identifiable Information". In: *Commun. ACM* 53 (2010), Nr. 6, S. 24–26. – ISSN 0001–0782

[Obe09] OBERSTE AUFSICHTSBEHÖRDE FÜR DEN DATENSCHUTZ IM NICHTÖFFENTLICHEN BEREICH: *Datenschutzkonforme Ausgestaltung von Analyseverfahren zur Reichweitenmessung bei Internet-Angeboten*. Stralsund, 2009

[Odl03] ODLYZKO, Andrew: Privacy, Economics, and Price Discrimination on the Internet. In: *ICEC '03: Proceedings of the 5th International Conference on Electronic Commerce*, ACM, 2003. – ISBN 1–58113–788–5, S. 355–366

[OGH05] OLSON, Judith S. ; GRUDIN, Jonathan ; HORVITZ, Eric: A Study of Preferences for Sharing and Privacy. In: *CHI '05: extended abstracts on Human Factors in Computing Systems*, ACM, 2005. – ISBN 1–59593–002–7, S. 1985–1988

[O'R05] O'REILLY, Tim: *What Is Web 2.0 – Design Patterns and Business Models for the Next Generation of Software*. 2005

[Org80] ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD): *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. 1980

[Org03] ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD): *Overview – Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. http://www.oecd.org/dataoecd/16/7/15589558.pdf Version: 2003

[Par95] PARLIAMENT, European: *Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data*. 1995

[Par97] PARLIAMENT, European: *Directive 97/66/EC on the processing of personal data and the protection of privacy in the telecommunications sector*. 1997

[Par00] PARLIAMENT, European: *Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market*. 2000

[Par02] PARLIAMENT, European: *Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector*. 2002

[PCT06] PASS, Greg ; CHOWDHURY, Abdur ; TORGESON, Cayley: A Picture of Search. In: *InfoScale '06: Proceedings of the 1st International Conference on Scalable Information Systems*, ACM, 2006. – ISBN 1–59593–428–6, S. 1

[PMTW08] PAPADOPOULOU, Elizabeth ; MCBURNEY, Sarah ; TAYLOR, Nick ; WILLIAMS, M. H.: Linking Privacy and User Preferences in the Identity Management for a Pervasive System. In: *WI-IAT '08: Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, 2008. – ISBN 978–0–7695–3496–1, S. 192–195

[Poi10] POINT TOPIC: *Broadband World Growth accelerates after the Doldrums of 2009*. http://point-topic.com/content/dslanalysis/BBAglobalbbq110.htm Version: 2010

[RDM09] REAY, Ian ; DICK, Scott ; MILLER, James: A large-scale empirical Study of P3P Privacy Policies: Stated Actions vs. Legal Obligations. In: *ACM Trans. Web* 3 (2009), Nr. 2, S. 1–34. – ISSN 1559–1131

[RJK08] REDDY, Madhu C. ; JANSEN, Bernhard J. ; KRISHNAPPA, Rashmi: The Role of Communication in Collaborative Information Searching. In: *Proceedings of the American Society for Information Science and Technology*, 2008

[RM03] RAO, Bharat ; MINAKAKIS, Louis: Evolution of mobile Location-based Services. In: *Commun. ACM* 46 (2003), Nr. 12, S. 61–65. – ISSN 0001–0782

[RMT+02] ROOS, T. ; MYLLYMÄKI, P. ; TIRRI, H. ; MISIKANGAS, P. ; SIEVÄNEN, J.: A probabilistic Approach to WLAN User Location Estimation. In: *International Journal of Wireless Information Networks* 9 (2002), Nr. 3, S. 155–164

[Ros07] ROSENBLUM, David: What Anyone Can Know: The Privacy Risks of Social Networking Sites. In: *IEEE Security and Privacy* 5 (2007), Nr. 3, S. 40–49

[Roß03] ROSSNAGEL, Alexander (. (Hrsg.): *Handbuch Datenschutzrecht: die neuen Grundlagen für Wirtschaft und Verwaltung*. München : Beck, 2003

[RRK05] RAFAELI, S. ; RABAN, D. ; KALMAN, Y.: Social Cognition Online. In: *The Social Net: Understanding Human Behavior in Cyberspace* (2005), S. 57–90

[Sav40] SAVIGNY, Friedrich C.: *System des heutigen Römischen Rechts*. Bd. 1. 1840

[SBB+05] SMYTH, Barry ; BALFE, Evelyn ; BOYDELL, Oisin ; BRADLEY, Keith ; BRIGGS, Peter ; COYLE, Maurice ; FREYNE, Jill: A Live-User Evaluation

of Collaborative Web Search. In: *IJCAI '05: Proceedings of the 19th International Joint Conference on Artificial intelligence*, Morgan Kaufmann Publishers Inc., 2005, S. 1419–1424

[SC+05] SMITH, Ian ; CONSOLVO, Sunny u. a.: Social Disclosure Of Place: From Location Technology to Communication Practice. In: *PERVASIVE '05: Proceedings of the 3rd International Conference on Pervasive Computing*, ACM, 2005, S. 81–90

[Ser10] SERVICES, Cross-Tab M.: *Online Reputation anlässlich des Europäischen Datenschutztages*. Microsoft, 2010

[SGB01] SPIEKERMANN, Sarah ; GROSSKLAGS, Jens ; BERENDT, Bettina: E-privacy in 2nd Generation E-Commerce: Privacy Preferences versus Actual Behavior. In: *Proceedings of the 3rd Conference on Electronic Commerce*, ACM Press, 2001. – ISBN 1–58113–387–1, S. 38–47

[Sok09] SOKOL, Bettina: Neunzehnter Datenschutz- und Informationsfreiheitsbericht. In: *Landesbeauftragte für Datenschutz und Informationsfreiheit Nordrhein-Westfalen*, 2009

[Spi07] SPIEKERMANN, Sarah: Privacy Enhancing Technologies for RFID in Retail- an Empirical Investigation. In: *UbiComp'07: Proceedings of the 9th International Conference on Ubiquitous Computing*. Berlin, Heidelberg : Springer-Verlag, 2007. – ISBN 978–3–540–74852–6, S. 56–72

[SSP09] SQUICCIARINI, Anna C. ; SHEHAB, Mohamed ; PACI, Federica: Collective Privacy Management in Social Networks. In: *WWW '09: Proceedings of the 18th International Conference on World Wide Web*, ACM, 2009. – ISBN 978–1–60558–487–4, S. 521–530

[Sta10] STATISTISCHE ÄMTER DES BUNDES UND DER LÄNDER: *Zensus 2011*. `http://www.statistik-portal.de/Statistik-Portal/zensus/` Version: 2010

[sti10] Datenschutz bei Onlinenetzwerken. In: *Test*, Stiftung Warentest, 4 2010

[Swe00] SWEENEY, L.: Uniqueness of Simple Demographics in the US Population. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh : LIDAP-WP4, 2000

[Swe02] SWEENEY, Latanya: k-Anonymity: A Model for Protecting Privacy. In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10 (2002), Nr. 5, S. 557–570. – ISSN 0218–4885

[TAJP07] TEEVAN, Jaime ; ADAR, Eytan ; JONES, Rosie ; POTTS, Michael A. S.: Information Re-Retrieval: Repeat Queries in Yahoo's Logs. In: *SIGIR '07: Proceedings of the 30th annual International Conference on Research and Development in Information Retrieval*, ACM, 2007. – ISBN 978–1–59593–597–7, S. 151–158

[TEG05] TINNEFELD, Marie-Theres ; EHMANN, Eugen ; GERLING, Rainer W.: *Einführung in das Datenschutzrecht : Datenschutz und Informationsfreiheit in europäischer Sicht*. 4., völlig neubearb. und erw. Aufl. Oldenbourg, 2005. – ISBN 3–486–27303–5

[The07] THE ASSOCIATED PRESS: Yahoo Criticized in Case of Jailed Dissident. In: *The New York Times* (2007), November 7. http://www.nytimes.com/2007/11/07/technology/07yahoo.html

[Tim08] TIMPF, Sabine: Location-based Services – Personalisierung mobiler Dienste durch Verortung. In: *Informatik-Spektrum* 31 (2008)

[TM08] TERROVITIS, Manolis ; MAMOULIS, Nikos: Privacy Preservation in the Publication of Trajectories. In: *MDM '08: Proceedings of the 9th International Conference on Mobile Data Management*, IEEE, 2008. – ISBN 978–0–7695–3154–0, S. 65–72

[TMK08] TERROVITIS, Manolis ; MAMOULIS, Nikos ; KALNIS, Panos: Privacy-preserving Anonymization of Set-Valued Data. In: *VLDB Endowment* 1 (2008), Nr. 1, S. 115–125. – ISSN 2150–8097

[TNP97] TWIDALE, Michael B. ; NICHOLS, David M. ; PAICE, Chris D.: Browsing is a collaborative Process. In: *Inf. Process. Manage.* 33 (1997), Nr. 6, S. 761–783. – ISSN 0306–4573

[TNS09] TNS INFRATEST BUSINESS INTELLIGENCE: 12. Faktenbericht, Bundesministerium für Wirtschaft und Technologie, 2009

[Una] UNABHÄNGIGES LANDESZENTRUM FÜR DATENSCHUTZ SCHLESWIG HOLSTEIN

[Una09] UNABHÄNGIGES LANDESZENTRUM FÜR DATENSCHUTZ SCHLESWIG-HOLSTEIN: *Datenschutzrechtliche Bewertung des Einsatzes von Google Analytics*. 2009

[Uni90] UNITED NATIONS (UN): *United Nations guidelines concering computerized personal data files*. http://www.worldlii.org/int/other/PrivLRes/1990/1.html, Version: 1990

[V+09] VYASA, Nitya H. u. a.: Towards Automatic Privacy Management in Web 2.0 with Semantic Analysis on Annotations. 2009. – Forschungsbericht

[Var07] VARIAN, H.R.: Position Auctions. In: *International Journal of Industrial Organization* 25 (2007), Nr. 6, S. 1163–1178

[Wes67] WESTIN, Alan F.: *Privacy and freedom*. London, 1967

[WJPT07] WIND, Rico ; JENSEN, Christian S. ; PEDERSEN, Kenneth H. ; TORP, Kristian: A Testbed for the Exploration of Novel Concepts in Mobile Service Delivery. In: *MDM '07: International Conference on Mobile Data Management*, IEEE, 2007. – ISBN 1–4244–1241–2, S. 218–220

[WLW98] WANG, Huaiqing ; LEE, Matthew K. O. ; WANG, Chen: Consumer Privacy Concerns about Internet Marketing. In: *Commun. ACM* 41 (1998), Nr. 3, S. 63–70. – ISSN 0001–0782

[WM07] WHITE, Ryen W. ; MORRIS, Dan: Investigating the Querying and Browsing behavior of Advanced Search Engine Users. In: *SIGIR '07: Proceedings of the 30th annual International Conference on Research and Development in Information Retrieval*, ACM, 2007. – ISBN 978–1–59593–597–7, S. 255–262

[Xam08] XAMIT BEWERTUNGSGESELLSCHAFT MBH: *Datenschutzbarometer*. http://www.xamit-leistungen.de/downloads/XamitDatenschutzbarometer2008.pdf. Version: 2008

[Xam09] XAMIT BEWERTUNGSGESELLSCHAFT MBH: *Datenschutzbarometer*. http://www.xamit-leistungen.de/downloads/XamitDatenschutzbarometer2009.pdf. Version: 2009

[XC07] XU, Toby ; CAI, Ying: Location Anonymity in Continuous Location-Based Services. In: *GIS '07: Proceedings of the 15th annual International Symposium on Advances in Geographic Information Systems*, ACM, 2007. – ISBN 978–1–59593–914–2, S. 1–8

[XT06] XIAO, Xiaokui ; TAO, Yufei: Anatomy: Simple and Effective Privacy Preservation. In: *VLDB '06: Proceedings of the 32nd International Conference on Very Large Data Bases*, VLDB Endowment, 2006, S. 139–150

[YW03] YU, Ting ; WINSLETT, Marianne: A Unified Scheme for Resource Protection in Automated Trust Negotiation. In: *SP '03: Proceedings of the Symposium on Security and Privacy*, IEEE, 2003. – ISBN 0–7695–1940–7, S. 110

[ZG09] ZHELEVA, Elena ; GETOOR, Lise: To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In: *WWW '09: 18th International World Wide Web Conference*, 2009, 531–531

[ZL08] ZOU, Deqing ; LIAO, Zhensong: A New Approach for Hiding Policy and Checking Policy Consistency, IEEE, 2008

F.2. Publikationsliste

[1] BURGHARDT, T. ; BÖHM, K. ; GUTTMANN, A. ; CLIFTON, C. : Search-Log Anonymization and Advertisement: Are They Mutually Exclusive? In: *CIKM '10: The 19th International Conference on Information and Knowledge Management*, ACM, 2010

[2] BURGHARDT, T. ; BUCHMANN, E. ; BÖHM, K. : Why do Privacy-Enhancement Mechanisms Fail, After All? A Survey of Both, the User and the Provider Perspective. In: *Web 2.0 and Trust*, 2008

[3] BURGHARDT, T. ; BUCHMANN, E. ; BOHM, K. : WI/IAT '08: Discovering the Scope of Privacy Needs in Collaborative Search. In: *International Conference on Web Intelligence and Intelligent Agent Technology* Bd. 1, IEEE, 2008. – ISBN 978–0–7695–3496–1, S. 910–913

[4] BURGHARDT, T. ; BUCHMANN, E. ; BÖHM, K. ; CLIFTON, C. : Collaborative Search And User Privacy: How Can They Be Reconciled? In: BERTINO, E. (Hrsg.) ; JOSHI, J. B. D. (Hrsg.): *Proceeding of the 4th International Conference on Collaborative Computing*, IEEE, 2008

[5] BURGHARDT, T. ; BUCHMANN, E. ; BÖHM, K. ; KÜHLING, J. ; BOHNEN, S. ; SIVRIDIS, A. : Tackling Compliance Deficits of Data-Protection Law with User Collaboration: A Feasibility Demonstration with Human Participants. In: *CEC '10: Conference on Commerce and Enterprise Computing*, IEEE, 2010

[6] BURGHARDT, T. ; BUCHMANN, E. ; BÖHM, K. ; KÜHLING, J. ; SIVRIDIS, A. : A Study on the Lack of Enforcement of Data Protection Acts. In: *e-Democracy '09: Next Generation Society, Technological and Legal Issues* Bd. 26, Springer, 2009. – ISBN 978–3–642–11629–2, S. 3–12

[7] BURGHARDT, T. ; BUCHMANN, E. ; MÜLLER, J. ; BÖHM, K. : Understanding User Preferences and Awareness: Privacy Mechanisms in Location-Based Services. In: *CoopIS '09: On the Move to Meaningful Internet Systems*, Springer, 2009

[8] BURGHARDT, T. ; WALTER, A. ; BUCHMANN, E. ; BOHM, K. : PRIMO - Towards Privacy Aware Image Sharing. In: *WI/IAT '08: International Conference on Web Intelligence and Intelligent Agent Technology (Workshops)* Bd. 3, IEEE, 2008. – ISBN 978–0–7695–3496–1, S. 21–24

F.3. Koautorenschaften

[9] KÜHLING, J. ; SIVRIDIS, A. ; SCHWUCHOW, M. ; BURGHARDT, T. : Das datenschutzrechtliche Vollzugsdefizit im Bereich der Telemedien – ein Schreckensbericht. In: *Springer, Datenschutz und Datensicherheit – DuD* (2008)

[10] WETH, C. von d. ; BOHM, K. ; BURGHARDT, T. ; HUTTER, C. ; YUE, J. Z.: Indirect Reciprocity in Policy-Based Helping Experiments. In: *ECOWS '09: European Conference on Web Services*, IEEE, 2009. – ISBN 978–0–7695–3854–9, S. 171–180