What is individual-related data?

Christian Eckert
Betreuer: Johann Schlamp
Seminar Innovative Internettechnologien und Mobilkommunikation WS12/13
Lehrstuhl Netzarchitekturen und Netzdienste
Fakultät für Informatik, Technische Universität München
Email: christian.eckert@in.tum.de

KURZFASSUNG

Der Schutz der Privatsphäre gewinnt durch die steigende Vernetzung und Nutzung des Internets an Bedeutung. In dieser Arbeit werden deshalb die rechtlichen Grundlagen in Deutschland und in der EU bezüglich personenbezogener Daten aufgearbeitet. Dabei stellt sich in der Praxis heraus, dass die Auslegung dieser Definitionen oftmals schwammig ist und zu kontroversen Gerichtsentscheidungen führt. In der Vergangenheit wurde in Forschungsprojekten gezeigt, dass es neben Cookies noch weitere Möglichkeiten gibt, einen Web-Browser über eine Art Fingerabdruck wiederzuerkennen. Dadurch ist es möglich, Bewegungsprofile einzelner Nutzer zu erstellen, was eine Bedrohung der Privatsphäre darstellt. Der Autor beschränkt sich hierbei auf den Informationsgehalt, der über den HTTP-Header des Web-Browsers gewonnen werden kann. Um dieser Bedrohung entgegenzuwirken, werden in dieser Arbeit verschiedene Maßnahmen und Modelle vorgestellt, die dem Schutz der Privatsphäre dienen. Dabei zeigt sich, dass es der Internetnutzer selbst in der Hand hat, dieses wichtige Gut zu bewahren.

Schlüsselworte

personenbezogene Daten, Datenschutz, Anonymität im Internet, HTTP-Header, k-anonymitiy, l-diversity, t-closeness

1. EINLEITUNG

Nicht nur durch die steigende Nutzung des Internets, sondern auch die Entwicklung hin zum Web 2.0 stellen eine Bedrohung der Privatsphäre im Internet dar. Dabei agiert der Benutzer im Internet nicht mehr nur als Konsument von Texten, Bildern und Videos, sondern wird dazu ermutigt, selbst eigene Inhalte zu veröffentlichen. Während diese Gefahr noch selbst reguliert und beispielsweise durch die Reduzierung der freiwilligen Preisgabe von persönlichen Inhalten gemindert werde kann, stellt die steigende Nutzung des Internets weiterhin ein Problem dar. Wie in einem Forschungsprojekt der Electronic Frontier Foundation demonstriert wurde, können Web-Browser nicht nur durch Cookies, sondern auch durch eine Reihe anderer Informationen identifiziert werden [9]. Die Auswertung dieses Projekts zeigt, dass die Kombination des zugrundeliegenden Betriebssystems, systemseitige Einstellungen wie Bildschirmauflösung und Spracheinstellungen, sowie die Wahl des Browsers und die darin installierten Plugins oftmals eine einzigartige Signatur ergeben, wodurch der Benutzer verfolgbar wird [8].

Mit diesem Hintergedanken werden in dieser Arbeit weniger die technischen Aspekte zur Erstellung eines Browser-Fingerabdrucks beleuchtet, sondern vielmehr die zugrundeliegenden gesetzlichen Regelungen zur Speicherung und zum Schutz von personenbezogenen Daten erläutert. In Kapitel 3 werden, basierend auf den gesetzlichen Bestimmungen, die Daten eines HTTP-Headers sowie die Log-Möglichkeiten eines Webservers näher betrachtet und bewertet. Verschiedene Modelle zum Schutz der Privatsphäre werden in Abschnitt 4 vorgestellt. Dabei werden mögliche Schutzmaßnahmen nicht nur aus der Perspektive des Benutzers vorgestellt, sondern auch auf die Rollen des Software-Entwicklers und des Webseiten-Betreibers eingegangen. Auf verwandte Arbeiten wird in Kapitel 5 verwiesen und Kapitel 6 enthält abschließend neben einer Zusammenfassung einen Ausblick zum Thema.

2. RECHTLICHE GRUNDLAGEN

Die nationalen und internationalen Beschlüsse und Gesetze rund um das Thema Datenschutz und Privatsphäre stellen die Grundlage dieser Arbeit dar. Darauf basierend können, die im Internet anfallenden Daten, beispielsweise die Inhalte eines HTTP-Header aus einem Webserver-Log, bezüglich ihrer rechtlichen Bedeutung, klassifiziert werden. Daraus wiederum resultieren die Vorschriften und die zu treffenden Maßnahmen für die Speicherung und den Schutz dieser Daten.

Einer der ersten Ansätze zum Schutz der Privatsphäre wurde im Jahre 1948 in der Generalversammlung der Vereinten Nationen beschlossen. Dabei kann das Recht auf Privatsphäre im weitesten Sinne als ein Bestandteil der international geltenden Allgemeinen Erklärung der Menschenrechte angesehen werden. Der Artikel 12 wurde dabei wie folgt definiert:

> "Niemand darf willkürlichen Eingriffen in sein Privatleben, seine Familie, seine Wohnung und seinen Schriftverkehr oder Beeinträchtigungen seiner Ehre und seines Rufes ausgesetzt werden. Jeder hat Anspruch auf rechtlichen Schutz gegen solche Eingriffe oder Beeinträchtigungen." [23]

Einen expliziten Schutz der Privatsphäre im Internet wurde dabei zwar nicht definiert - zumal das Internet damals noch gar nicht existierte - dennoch kann dies als erster globaler Grundstein zum Schutze des Privatlebens angesehen werden.

Aufbauend auf dieser Basis wurden sowohl in der Europäischen Union, als auch auf nationaler Ebene mit dem Bundes-

doi: 10.2312/NET-2013-02-1_12

datenschutzgesetz weitere Regelungen zum Schutz der Privatsphäre und folglich auch zum Schutz von personenbezogenen Daten im Internet definiert. Innerhalb der Europäischen Union ist dies in der Richtlinie 95/46/EG geregelt, die zum Schutz natürlicher Personen bei der Verarbeitung von personenbezogenen Daten und zur freien Datenverkehr dient [10].

Derzeit wird über eine Erneuerung der Datenschutzrichtlinie 95/46/EG verhandelt. Die ersten Reformvorschläge wurden bereits im Frühjahr 2012 präsentiert. Die Reform zielt im Besonderen darauf ab, ein einheitliches und hohes Datenschutzniveau in der EU sicherzustellen. Das Hauptaugenmerk liegt dabei vor allem auf den neuen technologischen Weiterentwicklungen wie beispielsweise mobiles Internet, Suchmaschinen und soziale Netzwerke. Das Ziel dabei ist die Sicherstellung der Transparenz in der Datenverarbeitung, die Gewährleistung der Betroffenenrechte sowie die Verpflichtung der Unternehmen, ihre Produkte von vornherein mit datenschutzfreundlichen Technologien auszustatten [11].

Die Umsetzung der Datenschutzrichtlinie 95/46/EG wird in Deutschland durch das Bundesdatenschutzgesetzt (BDSG) geregelt [1].

2.1 Was sind personenbezogene Daten?

Die zentrale Frage dieser Arbeit behandelt die Definition von personenbezogenen Daten, welche hier anhand der gesetzlichen Regelung in Deutschland als auch anhand der Richtlinie der Europäischen Union beantwortet werden soll. Innerhalb der EU Richtlinie steht der Ausdruck *personenbezogene Daten* für "[...] alle Informationen über eine bestimmte oder bestimmbare natürliche Person" (vgl. Artikel 2a Richtlinie 95/46/EG, [10]). Weiterhin wird eine Person als bestimmbar bezeichnet, wenn diese "direkt oder indirekt identifiziert werden kann, insbesondere durch Zuordnung zu einer Kennummer oder zu einem oder mehreren spezifischen Elementen, die Ausdruck ihrer physischen, physiologischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität sind" (vgl. Artikel 2a Richtlinie 95/46/EG, [10]).

An diese Richtlinie angelehnt wird in Deutschland folgende Begriffsbestimmung verwendet:

> "Personenbezogene Daten sind Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (Betroffener)." (vgl. § 3 Absatz 1 BDSG [1])

Einfache Beispiele personenbezogener Daten sind der vollständige Name oder die Anschrift einer Person. Mittels dieser Daten kann offensichtlich direkt eine bestimmte Person ausgemacht werden. Bei Informationen wie Familienstand, Alter oder Kontonummer handelt es sich zunächst um Einzelangaben über persönliche Verhältnisse. Auch der Besitz eines bestimmten Gegenstandes, beispielsweise eines Autos, ist als Einzelangabe sachlicher Verhältnisse einzuordnen. Diese Einzelangaben werden erst dann zu personenbezogenen Daten, wenn daraus eine bestimmte oder bestimmbare natürliche Person ableitbar ist. Dies ist immer dann der Fall, wenn diese Einzelangaben in einer Art Liste zusammen mit dem Namen gespeichert werden.

Deutlich schwieriger allerdings wird die Einstufung der Daten, wenn nicht direkt von einem einzelnen Datum auf eine Person geschlossen werden kann, aber aus einer Kombination verschiedener Einzelangaben. Bei einer Liste, die lediglich das Geschlecht und das Geburtsdatum verschiedener Personen enthält, kann sicherlich noch nicht von personenbezogenen Daten gesprochen werden. Was ist aber, wenn diese Liste neben dem Geschlecht und dem Geburtsdatum zusätzlich noch eine Postleitzahl enthält? Zunächst könnte angenommen werden, dass es sich selbst dann noch um eine anonyme Liste handelt. Die Informatikerin Latanya Sweeney hat jedoch gezeigt, dass allein durch diese Informationen 53% der Bevölkerung in den USA einzigartig und sogar 83% der US-Amerikaner mit großer Wahrscheinlichkeit identifiziert werden können [20].

Weiterhin spezifiziert das BDSG *besondere Arten* von personenbezogenen Daten. Dabei handelt es sich um die Angaben über religiöse oder philosophische Überzeugungen, die rassische und ethnische Herkunft, politische Meinungen, Gewerkschaftszugehörigkeit, Gesundheit oder Sexualleben (vgl. § 3 Absatz 9 BDSG [1]).

Neben den allgemeinen Bestimmungen im BDSG, die in § 1 bis § 11 definiert sind, wird zusätzlich zwischen der Datenverarbeitung seitens öffentlicher Stellen (§§ 12-26) und nicht-öffentlicher Stellen (§§ 27-38a) unterschieden. Unter nicht-öffentlichen Stellen sind (private) Unternehmen jeglicher Rechtsform aber auch öffentlich-rechtliche Wettbewerbsunternehmen zu verstehen. Abhängig vom Anwendungsbereich werden dabei die allgemeinen Bestimmungen aus § 1 bis § 11 bezüglich der Speicherung, Verarbeitung und Nutzung personenbezogener Daten im Einzelnen genauer definiert.

2.2 Verarbeitung und Schutz von personenbezogenen Daten

Bei der Zulässigkeit der Datenerhebung, -verarbeitung und -nutzung stimmen die Regelungen der EU sowie die Bestimmungen im BDSG größtenteils überein. Um personenbezogene Daten zu schützen, wird dabei eine Erhebung, Verarbeitung und Nutzung dieser Daten zunächst generell verboten. Ausnahmen werden nur dann erlaubt, wenn eine Einwilligung der betroffenen Person vorliegt oder eine andere Rechtsvorschrift dieses Gesetztes dies erlaubt (vgl. § 4 BDSG [1] und Artikel 7 Richtlinie 95/46/EG [10]). Zusätzlich muss nach § 28 des BDSG der Zweck der Datenverarbeitung oder -nutzung konkret festgelegt werden, wenn diese dem eigenen Geschäftszweck dient [1].

Weiterhin besagt § 3a des BDSG, dass die Systeme zur Erhebung, Verarbeitung und Nutzung personenbezogener Daten mit dem Ziel der Datenvermeidung und Datensparsamkeit auszurichten sind. Zudem müssen personenbezogene Daten anonymisiert oder pseudonymisiert werden, sofern dies nach dem Verwendungszweck möglich ist und nicht einen unverhältnismäßigen hohen Aufwand erfordert [1]. Eine solche Regelung ist in der Datenschutzrichtlinie 95/46/EG nicht enthalten.

doi: 10.2312/NET-2013-02-1_12

2.3 Regelung zur Speicherung von personenbezogenen Daten

Für elektronische Informations- und Kommunikationsdienste besagt das Telemediengesetz (TMG) nach § 13 Abs. 4 Nr. 2, dass anfallende personenbezogene Daten nach Ablauf der Nutzung unmittelbar gelöscht werden müssen. Die Verwendung von Nutzungsdaten über das Ende der Nutzungsdauer hinaus ist nach § 15 Abs. 4 TMG nur dann zulässig, wenn sie für die Zwecke der Abrechnung mit dem Nutzer erforderlich sind [2]. Darüber hinaus besagt § 9 des BDSG, dass bei der Verarbeitung und Speicherung von personenbezogene Daten ein Mindestmaß an Sicherheitsvorkehrungen getroffen werden muss. In der Anlage zu § 9 werden dabei konkrete Maßnahmen definiert. Dazu soll Unbefugten der Zutritt sowie der Zugriff auf die Datenverarbeitungssysteme verwehrt werden. Weiterhin muss bei der Übertragung von personenbezogenen Daten ein Verschlüsselungsverfahren verwendet werden, welchem dem Stand der Technik entspricht [1].

3. ANALYSE VON HTTP-HEADERN UND WEBSERVER-LOGS

Der HTTP-Header ist ein Bestandteil des Hypertext Transfer Protocol (HTTP), der wichtige Informationen bei der Kommunikation zwischen Browser und Webserver übermittelt. Hierbei soll sichergestellt werden, dass Nachrichten auf beiden Seiten korrekt interpretiert werden. Darüber hinaus wird die Möglichkeit geschaffen, mit Hilfe des HTTP-Headers die Webseite an die Gegebenheiten des Clients anzupassen. Dies betrifft beispielsweise die Anpassung der Sprache an die Voreinstellungen des Benutzers, oder auch die Optimierung der Webseite für mobile Geräte. Dabei wird unterschieden zwischen den Anfrage-Headerfeldern, welche vom Browser an den Webserver gesendet werden, und den Antwort-Headerfeldern, die der Webserver bei der Antwort an den Browser verwendet.

Dass der HTTP-Header nicht nur zur korrekten Interpretation der Daten sowie zur Optimierung der Webseite an den Client genutzt werden kann, hat das Forschungsprojekt der Electronic Frontier Foundation (EFF) gezeigt [9]. Dabei ist es gelungen, aus den erhaltenen Daten des HTTP-Headers sowie durch weitere Abfragen mittels Java, JavaScript und Flash einen Fingerabdruck des Browsers herzustellen. In dieser Arbeit jedoch liegt der Fokus ausschließlich auf dem HTTP-Header. Dabei ist in diesem Fall eine weitere Einschränkung auf die Anfrage-Headerfeldern möglich, da nur diese eine potenzielle Gefahr für den Benutzer darstellen. Die Header-Felder werden mittels Request for Comments (RFCs) von der Internet Engineering Task Force (IETF) spezifiziert. Zusätzlich dazu können Hersteller eigene Erweiterungen implementieren, die dann allerdings nicht von jedem Webserver bzw. Browser verstanden werden. Die aktuell gültige Spezifikation des Hypertext Transfer Protocol ist in RFC 2616 dokumentiert, welcher später durch weitere RFCs erweitert wurde [12].

3.1 Inhalt des HTTP-Headers

Mit Hilfe der gesetzlichen Bestimmungen aus dem vorherigen Kapitel ist eine Basis geschaffen, um die Inhalte eines HTTP-Headers bezüglich ihrer rechtlichen Relevanz einzuordnen. Zunächst soll Tabelle 1 einen Überblick der in einem HTTP-Header enthaltenen Felder bieten. Dabei handelt es sich lediglich um einen Auszug der Anfrage-Felder, die im Rahmen dieser Arbeit sinnvoll erscheinen.

Tabelle 1: Auszug der HTTP-Header Anfrage-Felder

Feld	Beschreibung
Accept	Vom Browser akzeptierte Dateitypen
Accept-Charset	Vom Browser unterstützte Zeichensätze
Accept-Encoding	Gibt die unterstützten komprimierten Formate an
Accept-Language	Unterstützte Spracheinstellungen
Authorization	Enthält Authentifizierungsdaten
Expect	Besschreibt das Verhalten, das der Client vom Server erwartet
From	Beinhaltet die E-Mail Adresse des Nutzers; allerdings muss der Nutzer ausdrücklich zustimmen, wenn diese gesendet wird
Host	Domain-Name des Webservers
Referer	Enthält die URI der Webseite von der aus auf die aktuelle Webseite über ein Link verwiesen wurde
User-Agent	Beinhaltet Informationen über den Client, wie zum Beispiel Betriebssystem und Webbrowser

Ein weiteres für diese Arbeit relevantes Anfrage-Headerfeld wurde in RFC 4229 hinzugefügt. Dabei handelt es sich um das sogenannte Cookie, das von dem Webbrowser gespeichert wird und dazu dient, einen Browser beim erneuten Besuchen einer Webseite wiederzuerkennen [18].

3.2 Rechtliche Einstufung der Daten

Die in Tabelle 1 aufgelisteten Anfrage-Headerfelder sind einzeln für sich betrachtet aus gesetzlichen Gesichtspunkten eher unkritisch, da es sich weder um sensible Informationen noch um Schlüsselattribute handelt, mit Hilfe derer eine Person direkt identifiziert werden kann. Einzige Ausnahme stellt das *From* Header-Feld dar, welches die E-Mail Adresse des Benutzers enthält. Allerdings wurde in RFC 2616 beschrieben, dass diese Information nicht ohne die Einwilligung des Nutzers gesendet werden sollte, da dies mit den Datenschutzinteressen des Nutzers in Konflikt stehen könnte. Des Weiteren sollte der Benutzer die Möglichkeit haben, das Feld zu deaktivieren bzw. den Inhalt des Feldes vor dem Absenden der Anfrage zu modifizieren [12]. Auch wenn die E-Mail-Adresse definitiv als personenbezogenes Datum einzustufen ist, ist das *From* Header-Feld dennoch konform mit den gesetzlichen Regelungen, da eine ausdrückliche Einwilligung des Nutzers vorhergehen muss (vgl. § 4 BDSG [1] und Artikel 7 Richtlinie 95/46/EG [10]). Eine mögliche Problematik diesbezüglich besteht auch deswegen nicht, da die Verwendung dieses Header-Feldes heutzutage kaum Verwendung findet. Offen allerdings bleibt die Frage, ob die Kombination sämtlicher Headerfelder so aussagekräftig ist, um eine Person eindeutig zu identifizieren. Auf diese Frage wird im folgenden Kapitel näher eingegangen.

Neben diesen HTTP-Header-Feldern kann ein Webserver weitere Informationen aus einem TCP/IP Datenpakt extrahieren. Hierbei ist vor allem die IP-Adresse zu nennen. Inwie-

doi: 10.2312/NET-2013-02-1_12

fern die IP-Adresse als personenbezogenes Datum angesehen wird, ist aktuell noch umstritten. Datenschützer argumentieren, dass bei einer statisch vergebenen IP-Adresse ein direkter und andauernder Bezug zum Anschlussinhaber möglich ist. Doch auch selbst bei einer dynamischen zugewiesenen IP-Adresse in Kombination mit dem Zeitstempel des Zugriffs wird argumentiert, dass in vielen Fällen mit Hilfe Dritter ein Bezug zu einer bestimmbaren Person hergeleitet werden kann. Diese Auffassung wurde in verschiedenen Gerichtsentscheidungen [14][4], in einem öffentlichen Schreiben des Bundesjustizministerium [7] sowie vom Bundesamt für Sicherheit in der Informationstechnik (BSI) [6] geteilt. Allerdings liegen auch Gerichtsbeschlüsse vor, bei denen die IP-Adresse nicht als personenbezogenes Datum aufgefasst wurde [3]. Ein eindeutiger und gesetzlich haltbarer Beschluss fehlt damit weiterhin. Durch den Vorschlag des Europäischen Parlaments zur Reform der 95/46/EG ist allerdings die Tendenz erkennbar, wonach zukünftig IP-Adressen als personenbezogenes Datum eingestuft werden sollen [11].

3.3 Bewertung des Informationsgehalts

Um die verschiedenen Variablen bezüglich ihres Informationsgehaltes vergleichen zu können, wird die Entropie als Maß der Informationsdichte verwendet. Dabei entspricht die Entropie, die in Bits gemessen wird, dem mittleren Informationsgehalt einer Messvariable. Der Informationsgehalt wiederum hängt von der Auftrittswahrscheinlichkeit der Messvariable ab. Generell gilt, je geringer die Auftrittswahrscheinlichkeit einer Variable ist, desto höher ist der Informationsgehalt und andersherum. Wichtig in diesem Zusammenhang ist zu erwähnen, dass die Entropie verschiedener Variablen nur dann miteinander addiert werden kann, sofern keine statistische Abhängigkeit zwischen den Messvariablen besteht [8].

Die folgenden Informationen beziehen sich auf die Auswertung des Forschungsprojekts Panopticlick der EFF. Die Analyse der Daten basiert auf einem Datensatz von insgesamt über einer Million Klicks auf ihrer Test-Webseite, wovon durch eine Reduktion von redundanten Daten immer noch 470.161 Datensätze übrig geblieben sind.

Tabelle 2: Durchschnittliche Entropie der HTTP-Felder [8, S. 5-17]

Header-Feld	Entropie	Quelle
User-Agent	10,0	HTTP-Header
Accept-Headers	6,09	HTTP-Header
Cookies erlaubt?	0,353	HTTP-Header
Schriftarten	13,9	Flash / Java Applet
Browser Plugins	15,4	JavaScript / AJAX
Bildschirm-Auflösung	4,83	JavaScript / AJAX
Zeitzone	3,04	JavaScript / AJAX
Supercookie Test	2,12	JavaScript / AJAX

Für den *User-Agent* bedeutet dies beispielsweise, dass wenn aus der Gesamtmenge der Datensätze ein Browser zufällig ausgewählt wird, nur jeder 2^{10}-te (=1024) Browser genau die gleiche *User-Agent* Charakteristik aufweist [8]. Der *User-Agent* ist somit ein relativ aussagekräftiges Kriterium bei der Identifikation von Browsern. Neben den gewonnen Informationen über den HTTP-Header sind in dieser Tabelle noch weitere Kriterien aufgelistet, die über Java, Javascript (bzw. AJAX) und Flash gewonnen werden, damit die Werte besser eingeordnet werden können.

Leider enthält die statistische Auswertung von Eckersley keine Aufschlüsselung der einzelnen Accept-Header-Feldern. Stattdessen sind die vier in Tabelle 1 beschriebenen Accept-Header zusammengefasst worden, weshalb eine Aussage zu den einzelnen Feldern nicht möglich ist. Dennoch lässt sich ablesen, dass die Accept-Felder zwar einen nicht zu vernachlässigbaren Anteil zur Identifikation der Browser ausmachen, aber in der Gesamtsumme eher untergehen. Kaum Bedeutung hingegen ist der Abfrage, ob Cookies erlaubt sind oder nicht, zuzumessen. Hierbei ist anzumerken, dass die einzelnen Header-Felder statistische Abhängigkeiten besitzen. Die Abfrage, ob Cookies erlaubt sind oder nicht, ist beispielsweise von den Grundeinstellungen des Web-Browser und somit vom User-Agent abhängig, weshalb die Entropie-Bits nicht miteinander addiert werden dürfen.

Des Weiteren ist aus der Tabelle abzulesen, dass generell die Aussagekraft, die über den HTTP-Header gewonnen werden kann, eher gering ist im Vergleich zu den Möglichkeiten, die JavaScript bzw. AJAX bietet. Allerdings muss an dieser Stelle relativiert werden, dass der HTTP-Header tatsächlich bei jeder Anfrage übertragen wird und nur schwer zu blocken ist. JavaScript hingegen kann in den gängigen Browsern entweder direkt, oder durch diverse Plugins deaktiviert werden. Es ist allerdings auch schon für die Browser Firefox [17] und Chrome [13] ein Plugin vorhanden, welches die Modifikation des HTTP-Anfrage-Headers ermöglicht. Laut Beschreibung sind diese Plugins allerdings primär an Web-Entwickler gerichtet, um die Funktionalität einer Webseite zu testen, wenngleich auch das Thema Schutz der Privatsphäre erwähnt wird.

Welcher Informationsgehalt der IP-Adresse in diesem Zusammenhang beizumessen ist, geht aus der Arbeit von Eckersley nicht hervor. Dennoch ist festzuhalten, dass es sich bei der IP-Adresse um eine Variable handelt, die sich im Regelfall täglich verändert. Bei den bereits besprochenen Messvariablen sind die Werte hingegen über einen längeren Zeitraum relativ stabil. Diese Eigenschaft ist auch bei der Erstellung eines Browser-Fingerabdrucks erforderlich. Durch die Verwendung von Network Address Translation (NAT) unter IPv4 wird die Zuordnung der IP-Adresse zu einem Browser weiter erschwert. Eine fundierte Aussage über den Informationsgehalt der IP-Adresse ist daher nur schwer möglich. Eine statische und fest zugewiesene IP-Adresse würde hingegen, in Verbindung mit den anderen Messvariablen, die Wiedererkennung der Geräte aus diesem Internetanschluss deutlich vereinfachen.

4. MAßNAHMEN UND MODELLE ZUM SCHUTZ DER PRIVATSPHÄRE

Um ein Mindestmaß an Schutz der Privatsphäre eines Internetnutzers zu gewährleisten, existieren verschiedene Ansätze. Dabei hängen die Beweggründe als auch die Möglichkeiten stark von der jeweiligen Rolle ab. Im Folgenden werden dazu die Perspektiven des Benutzers, des Software-Entwicklers als auch die eines Webseiten-Betreibers näher betrachtet.

doi: 10.2312/NET-2013-02-1_12

4.1 Aus Sicht eines Internetnutzers

Wem der Schutz der eigenen Privatsphäre wichtig ist, kann selbst dazu beitragen, dass dieser Schutz gewährleistet ist. Dazu bieten alle gängigen Web-Browser verschiedenste Einstellmöglichkeiten an. So kann beispielsweise konfiguriert werden, ob Cookies überhaupt erlaubt und wenn ja, nach welchem Zeitpunkt wieder gelöscht werden sollen. Ebenso kann JavaScript komplett deaktiviert werden. Deutlich komfortabler und flexibler können diese Einstellungen auch über verschiedene Plugins vorgenommen werden, welche bei den meisten Browsern nachinstalliert werden können. Um zusätzlich noch die eigene IP-Adresse zu verschleiern, kann die Verbindung über einen ausländischen Proxy geleitet werden. Noch weitreichender sind die Möglichkeiten, die ein Anonymisierungs-Tool wie beispielsweise Tor bietet [5]. Bei der großen Vielzahl an Plugins gilt es allerdings zu beachten, dass einige Schutzmaßnahmen auch einen kontraproduktiven Einfluss auf die Identifizierbarkeit des Browser haben können. Dies ist vor allem dann der Fall, wenn ein Plugin installiert wurde, das nur von einer geringen Anzahl an Nutzern verwendet wird, wodurch sich der Fingerabdruck des Browsers auf einen kleinen Nutzerkreis eingrenzen lässt [8].

Der Kompromiss, den man mit solchen Schutzmaßnahmen eingeht, ist meist ein Verlust an Komfort. Im Regelfall wird durch die genannten Maßnahmen das Surf-Erlebnis reduziert, da einzelne Webseiten nicht korrekt dargestellt werden können, oder die Latenz der Verbindung erhöht wird.

An dieser Stelle ebenso zu nennen, ist ein verantwortungsvoller Umgang mit den eigenen Daten, die man selbst veröffentlicht. Denn auch die besten Anonymisierungs-Tools und Plugins sind nutzlos, wenn im Internet freiwillig intime Daten preisgegeben werden, wie beispielsweise in Sozialen Medien.

4.2 Maßnahmen für Software-Entwickler

Aus der Perspektive eines Software-Herstellers, egal ob es sich dabei um die Software eines Browsers oder die einer Browser-Komponente handelt, mag es zunächst wenig Beweggründe geben, sich um den Schutz der Privatsphäre zu kümmern; schließlich resultiert daraus kein direkter Nutzen. In diesem Zusammenhang ist vor allem die Übertragung von detaillierten Versionsnummern in den HTTP-Header-Feldern gemeint, die dann laut Eckersley bei der Erstellung eines Device-Fingerprints zu Lasten der Privatsphäre genutzt werden können [8]. Die Software-Entwickler selbst hingegen können von der Angabe dieser Mikroversionen profitieren, beispielsweise bei der Analyse von Fehlverhalten einzelner Versionen oder auch bei der Erstellung einer Statistik bezüglich der Nutzungsverteilung der einzelnen Versionen der Software.

Dennoch gibt es auch Gründe für einen Software-Hersteller, bei der Entwicklung der Produkte auf Datenschutzvorkehrungen zu achten. Wenn beispielsweise durch verschiedenste Maßnahmen die Privatsphäre der Benutzer besonders geschützt wird, könnte dies das Vertrauen der Benutzer in die Produkte stärken, wodurch die Reputation des Unternehmens steigt, was wiederum zu steigenden Nutzungszahlen führen kann. Ebenso ist es denkbar, dass Software-Entwickler gesetzlich aufgefordert werden, die Privatsphäre der Nutzer von vornherein zu schützen. Dieses Szenario könnte sogar schon bald durch eine mögliche Reform der Richtlinie 95/46/EG Realität werden [11].

4.3 Modelle zur Datenspeicherung seitens des Webseiten-Betreibers

Wenn einem Webseiten-Betreiber die Privatsphäre der Nutzer am Herzen liegt, dann gibt es eine simple Methode dies zu realisieren: Indem gar keine Daten gespeichert werden. Dies beinhaltet den Verzicht der Nutzung von Cookies, Logging und im Idealfall auch eine Vermeidung von JavaScript. Derzeit lässt sich solch einen Zusammenschluss mehrerer Webseiten finden, die dem Benutzer versprechen, keine personenbezogene Daten zu speichern. Ebenso dürfen auch keine externen Dienste wie Statistiken oder Werbung eingebunden werden, die diese Regeln missachten. Jeder Webseiten-Betreiber, der an dieser Aktion teilnehmen will und die Bedingungen erfüllt, darf schlussendlich ein Gütesiegel auf der eigenen Webseite einbinden [24].

Dennoch lässt es sich in manchen Situationen nicht vermeiden, auch personenbezogene Daten zu speichern. Dies ist beispielsweise dann der Fall, wenn die Daten - im Einklang mit den gesetzlichen Bestimmungen - zu Zwecken der Abrechnung über die Nutzungsdauer hinaus gespeichert werden müssen. Weiterhin fordert das Gesetz eine Anonymisierung oder Pseudonymisierung der Daten, sofern dies nicht mit einem unverhältnismäßig hohen Aufwand verbunden ist.

Zur Veranschaulichung werden an dem Beispiel einer medizinischen Tabelle verschiedene Modelle zur Datenspeicherung vorgestellt.

Tabelle 3: Medizinische Krankheitstabelle

Name	Geburts-datum	Ge-schlecht	PLZ	Krankheit
Hans Maier	05.03.82	M	81247	Diabetes
Max Müller	13.08.82	M	81252	Diabetes
Ida Schmitt	01.08.81	W	81246	Tumor
Anna Meier	12.01.82	W	81247	HIV

Das sensitive Attribut in dieser Tabelle ist die Krankheit. So wie die Tabelle hier vorliegt, steht die Erkrankung in einem direkten Bezug zu einer Person. Aber auch eine Anonymisierung, bei der lediglich die Namen aus dieser Tabelle gestrichen werden, ist nicht ausreichend. Denn auch die Attribute Geburtsdatum, Geschlecht und PLZ sind in diesem Beispiel ausreichend, um eine Personen eindeutig identifizieren zu können. Eine solche Kombination von Attributen wird auch als *quasi-identifier* bezeichnet [21]. Der Name hingegen ist ein Schlüsselattribut, da über diesen eine Person direkt identifiziert werden kann.

4.3.1 k-anonymity

Um eine solche indirekte Identifikation mittels *quasi-identifier* zu verhindern, kann das Modell *k-anonymity* eingesetzt werden. Dieses Modell besagt, dass für jeden Datensatz und für jede beliebige Kombination aus *quasi-identifiern* mindestens k-1 andere ununterscheidbare Datensätze existieren müssen [15].

Mit dem Wissen über Geburtsdatum, Geschlecht und PLZ einer einzelnen Person ist es mit den Informationen aus Ta-

doi: 10.2312/NET-2013-02-1_12

Tabelle 4: Anonymisierte Tabelle mit k=2

Geburts-datum	Ge-schlecht	PLZ	Krankheit
..82	M	812**	Diabetes
..82	M	812**	Diabetes
..8*	W	8124*	Tumor
..8*	W	8124*	HIV

belle 4 generell nicht mehr möglich, auf die Krankheit einer Person zu schließen, da mindestens zwei Personen in Frage kommen. Je höher der Wert von k ist, desto besser ist die Anonymisierung eines Individuums. Mathematisch ausgedrückt bedeutet dies, dass durch die Kenntnis eines *quasi-identifier* eine ausgewählte Person nur mit einer Wahrscheinlichkeit von $\frac{1}{k}$ bestimmt werden kann. Allerdings existieren verschieden Angriffsmöglichkeiten gegen dieses Modell. Beispielsweise kann mit dem Wissen, dass die Person Hans Maier in dieser Tabelle gelistet ist und Geburtsdatum, PLZ und Geschlecht bekannt sind, aus Tabelle 4 abgelesen werden, dass Hans Maier an Diabetes leidet, da dies für beide in Frage kommenden Personen gilt [15].

4.3.2 ℓ-diversity

Um genau die beschriebene Schwachstelle von *k-anonymity* anzugehen, wurde das Modell *ℓ-diversity* eingeführt, das einen besseren Schutz bietet. Dabei wird zunächst zwischen sensitiven und nicht-sensitiven Attributen unterschieden. Weiterhin muss neben den Bedingungen von *k-anonymity* zusätzlich noch gelten, dass für jede mögliche Kombination mindestens ℓ-verschiedene sensitive Attribute vorhanden sind. Wenn Hans Maier beispielsweise statt Diabetes eine andere Krankheit hätte, dann würde Tabelle 4 die Eigenschaften von *ℓ-diversity* (mit ℓ=2) erfüllen. Allerdings stößt auch dieses Modell an seine Grenzen, wenn die sensitiven Attribute in den Datensätzen unausgeglichen sind oder auch wenn ein sensitives Attribut lediglich mit Ja/Nein beantwortet werden kann. Unter diesen Umständen kann der Datensatz im besten Fall lediglich die Bedingungen für ℓ=2 erfüllen [16].

4.3.3 t-closeness

Das Konzept *t-closeness* ist ein weiteres Modell zur Anonymisierung von Daten, das auf *ℓ-diversity* aufbaut. Zunächst bezeichnen wir eine Menge an Tupeln, die sich durch einen *quasi-identifier* nicht unterscheiden lassen, als *q*-Block*. Dabei wird versucht, den Gewinn an Wissen durch die anonymisierten Daten möglichst gering zu halten. Dies wird in diesem Modell dadurch sichergestellt, indem die Wahrscheinlichkeitsverteilung der sensitiven Attribute jedes *q*-Blocks* nur eine maximale Distanz t zu der Verteilung der sensitiven Attribute in der gesamten Tabelle enthält. Die Distanz zweier Wahrscheinlichkeitsverteilungen kann wiederum mathematisch berechnet werden, beispielsweise mit Hilfe der Kullback-Leibler-Divergenz. Ziel dieses Modells ist es, den Wert von t möglichst gering zu halten. Die gesamte Tabelle besitzt genau dann die Eigenschaft *t-closeness*, wenn für jeden *q*-Block* die Eigenschaft *t-closeness* erfüllt ist [15]. Konkret bedeutet dies also, dass der Wert von t genau dann klein ist, wenn die prozentuale Verteilung der sensitiven Attribute für jeden *q*-Block* möglichst identisch ist.

Wenn die Verteilung der sensitiven Attribute innerhalb der *q*-Blöcke* hingegen sehr verschieden ist, könnte wie folgt ein Angriff durchgeführt werden. Angenommen eine Tabelle enthält ein sensitives Attribut, das die Information beinhaltet, ob eine Person an HIV erkrankt ist oder nicht. Dabei seien 99% der Personen gesund und 1% habe die Krankheit HIV. In einem der *q*-Blöcke* allerdings ist die Verteilung der gesunden und erkrankten Personen ausgeglichen, also im Verhältnis von 50:50. Wenn die *quasi-identifier* einer gesuchten Person auf diesen *q*-Block* zutreffen, dann kann daraus geschlossen werden, dass die gesuchte Person mit einer Wahrscheinlichkeit von 50% HIV hat. Damit erhält ein potenzieller Angreifer einen nicht vernachlässigbaren Informationsgewinn über eine einzelne Personen gegenüber der Allgemeinheit. Würde die Tabelle hingegen die Eigenschaft *t-closeness* mit einem kleinen Wert für t erfüllen, wäre ein solcher Angriff nicht möglich beziehungsweise der Wissensgewinn wäre deutlich geringer, da ein solch abweichender *q*-Block* gar nicht vorkommen dürfte.

5. VERWANDTE ARBEITEN

Während in dieser Arbeit der Fokus auf den rechtlichen Bestimmungen zum Thema Datenschutz liegt, wurden in der Seminararbeit von Thomas Pieronczyk die technischen Hintergründe zur Erstellung der Browser-Fingerabdrücke näher erläutert [19]. Diese Arbeit basierte ebenfalls auf den statistischen Auswertungen von Eckersley [8], in der eine ausführliche Beschreibung der Datensammlung sowie eine Deutung der Ergebnisse enthalten ist.

Aus aktuellem Anlass ebenfalls sehr interessant ist in diesem Zusammenhang das Projekt von Henning Tillmann, der im Rahmen seiner Diplomarbeit selbst Untersuchungen zur Identifizierbarkeit von Web-Browsern durchführt. Die Auswertung des Projekts lag bei der Entstehung dieser Arbeit allerdings noch nicht vor, soll aber im Frühjahr 2013 veröffentlicht werden [22].

Eine ausführliche Erläuterung, der in dieser Arbeit vorgestellten Modelle zur Anonymisierung der Daten, wird in dem Artikel von Latanya Sweeney [21] sowie in Referenz [15] sehr gut beschrieben und mit zahlreichen Beispielen veranschaulicht. Dabei werden auch die mathematischen Hintergründe ausgiebig erklärt.

6. ZUSAMMENFASSUNG UND AUSBLICK

Trotz der expliziten Definition von personenbezogenen Daten in den Gesetzesbüchern bestehen nach wie vor kontroverse Meinungen bezüglich der Auslegung dieser Definition. Die größten Differenzen beziehen sich auf die Fragestellung, inwiefern aus den übertragenen Daten eine bestimmte oder bestimmbare Person ableitbar ist.

Die Analyse der in dem HTTP-Header enthaltenen Feldern hat gezeigt, dass diese vielmehr nützlich als schädlich und in der Praxis alleine kaum ausreichend sind, um ein effektives Device-Fingerprinting zu ermöglichen. Deutlich kritischer und umfangreicher ist der Informationsgehalt der mittels Java, Flash und JavaScript gewonnen werden kann, wie Eckersley gezeigt hat [8]. Ebenfalls als bedenklich ist die IP-Adresse einzustufen. Denn mit Hilfe dieser ist prinzipiell die Möglichkeit gegeben, eine Verbindung zwischen den Daten und einer bestimmbaren Person herzustellen.

doi: 10.2312/NET-2013-02-1_12

Die verschiedenen Maßnahmen zum Schutz der Privatsphäre haben verdeutlicht, dass auf Seiten des Benutzers am meisten Potenzial vorhanden ist, um die eigene Privatsphäre zu schützen – auch wenn dies meist mit einem Verlust an Komfort verbunden ist. Allerdings gilt zu beachten, dass einige Maßnahmen auch einen kontraproduktiven Effekt haben können. Ebenfalls interessant sind die vorgestellten Modelle, die einem Web-Hoster zur Verfügung stehen, um bei der Speicherung von personenbezogenen Daten eine bestmögliche Anonymisierung zu ermöglichen.

Mit Spannung zu erwarten ist die Entwicklung der gesetzlichen Beschlüsse, denn nur wenn dieser Grundstein gelegt ist, kann darauf aufbauend ein erhöhtes Maß an Privatsphäre im Internet sichergestellt werden. Dazu müssen aber zunächst die Menschen bezüglich der Bedeutung der Privatsphäre sensibilisiert werden. Denn nur wenn der Wunsch und die Forderung seitens der Bürger zu mehr Datenschutz besteht, ist eine Reaktion der Legislative denkbar.

7. LITERATUR

[1] Bundesdatenschutzgesetz in der Fassung der Bekanntmachung vom 14. Januar 2003 (BGBl. I S. 66), das zuletzt durch Artikel 1 des Gesetzes vom 14. August 2009 (BGBl. I S. 2814) geändert worden ist. http://www.gesetze-im-internet.de/bundesrecht/bdsg_1990/gesamt.pdf, Dec. 1990.

[2] Telemediengesetz vom 26. Februar 2007 (BGBl. I S. 179), das zuletzt durch Artikel 1 des Gesetzes vom 31. Mai 2010 (BGBl. I S. 692) geeändert worden ist. http://www.gesetze-im-internet.de/bundesrecht/tmg/gesamt.pdf, Feb. 2007.

[3] Amtsgericht München. Urt. vom 30.09.2008 - Az. 133 C5677/08. http://tlmd.in/u/524.

[4] Amtsgericht Wuppertal. Urt. vom 03.04.2007 - 29 Ds 70 Js 6906/06. http://www.jurpc.de/jurpc/show?id=20080110.

[5] J. Appelbaum. The Tor Project. https://www.torproject.org/about/overview.html.en. Abgerufen am 8 Februar 2013.

[6] Bundesamt für Sicherheit in der Informationstechnik. Datenschutzgerechtes E-Government. In *E-Government-Handbuch*, pages 12–19. Bundesanzeiger, Köln, 2005.

[7] Bundesministerium der Justiz. R B 3 - zu 4104/8 - 1 -R5 39/2008. http://www.daten-speicherung.de/data/bmj_2009-02-02.pdf.

[8] P. Eckersley. How unique is your web browser? In *Proceedings of the 10th international conference on Privacy enhancing technologies*, PETS'10, pages 1–18, Berlin, Heidelberg, 2010. Springer-Verlag.

[9] Electronic Frontier Foundation. https://panopticlick.eff.org/. Abgerufen am 05 Januar 2013.

[10] EU-Parlament. Richtlinie 95/46/EG. *Amtsblatt*, (L 281):31–50, Oct. 1995. 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.

[11] Europäische Kommision. Vorschlag für Verordnung des Europäischen Parlaments und des Rates zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr (Datenschutz-Grundverordnung), Jan. 2012.

[12] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1. RFC 2616, IETF, June 1999.

[13] Google Chrome. Change HTTP Request Header. https://chrome.google.com/webstore/detail/change-http-request-heade/ppmibgfeefcglejjlpeihfdimbkfbbnm. Abgerufen am 10 Februar 2013.

[14] Landgericht Berlin. Urt. vom 06.09.2007 - 23 S 3/07, MMR 2007, 799-800. http://www.daten-speicherung.de/data/Urteil_IP-Speicherung_2007-09-06.pdf.

[15] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, Apr. 2007.

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *TKDD*, 1(1), 2007.

[17] Mozilla Corporation. Modify Headers. https://addons.mozilla.org/en-us/firefox/addon/modify-headers/. Abgerufen am 15 Dezember 2012.

[18] N. Nottingham and J. Mogul. HTTP Header Field Registrations. RFC 4229, IETF, Dec. 2005.

[19] T. Pieronczyk. Device Fingerprinting mit dem Web-Browser. In G. Carle and C. Schmitt, editors, *Proceedings of the Seminars Future Internet (FI), Innovative Internet Technologies and Mobile Communication (IITM) and Aerospace Networks (AN)*, volume NET-2012-08-1, pages 23–29, Aug. 2012.

[20] L. Sweeney. Uniqueness of Simple Demographics in the U.S. Population, 2000.

[21] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, Oct. 2002.

[22] H. Tillmann. Browser Fingerprinting. http://bfp.henning-tillmann.de/. Abgerufen am 5 Januar 2013.

[23] United Nations. The Universal Declaration of Human Rights. 1948.

[24] Wir speichern nicht! http://www.wirspeichernnicht.de/. Abgerufen am 19. Dezember 2012.

doi: 10.2312/NET-2013-02-1_12