What is "Privacy"? - Information theory

Samuel Hall
Betreuer: Marcel von Maltitz
Seminar Innovative Internet-Technologien und Mobilkommunikation SS2015
Lehrstuhl Netzarchitekturen und Netzdienste
Fakultät für Informatik, Technische Universität München
Email: sammy.hall@tum.de

KURZFASSUNG

Wenn man versucht eine Definition für Datenschutz zu finden, stößt man sehr schnell auf das Problem, dass es sehr viele verschiedene Definition gibt, die sich zum einen stark voneinander unterscheiden und zum anderen oft sehr weich formuliert sind. In dieser Seminararbeit werden deshalb sowohl verschieden Metriken vorgestellt, die sich an verschiedene Definitionen von Datenschutz und seinen Teilaspekte anlehnen, als auch auf die Probleme eingegangen, die dabei entstehen. Ziel soll sein Datenschutz verschiedener Systeme mit Hilfe von diesen Metriken abzuprüfen. Dabei soll vor allem der informations-theoretische Ansatz im Vordergrund stehen.

Schlüsselworte

Privacy, Privatsphäre, Datenschutz, Metrik, Informationstheorie, anonymity, unlinkability

1. EINLEITUNG

Um *"What is privacy?"* beantworten zu können, muss zunächst für den Begriff *privacy* eine korrekte Übersetzung ins Deutsche gefunden werden. Vergleicht man verschiedene Übersetzungen [1][2], stellt man fest, dass sich *privacy*, sowohl zu Privatsphäre, als auch zu Datenschutz übersetzen lässt. Die zwei Begriffe sind aber keinesfalls als Synonyme zu verstehen. Datenschutz lässt sich auf vielfältige Weise definieren. Der Duden definiert Datenschutz folgendermaßen:

> Schutz des Bürgers vor Beeinträchtigungen seiner Privatsphäre durch unbefugte Erhebung, Speicherung und Weitergabe von Daten, die seine Person betreffen. [3]

Datenschutz ist hier also als Schutz der Privatsphäre definiert. Privatsphäre selbst lässt sich aus Artikel 12 der Allgemeinen Erklärung der Menschenrechte [4] ableiten und ist deshalb auch im deutschem Recht in Form verschiedener Gesetze, wie dem Recht auf Post- und Fernmeldegeheimnis oder das Recht auf die Unverletzlichkeit der Wohnung im Grundgesetz [5], verankert. Durch diese Gesetze und ihrer tiefen Verankerung wird deutlich wie wichtig Privatsphäre für die Gesellschaft ist. Datenschutz trägt maßgeblich zum Erhalt der Privatsphäre bei. Datenschutz selbst wird im Grundgesetz allerdings nicht explizit erwähnt. Wenn man sich die Duden Definition anschaut, könnte man auch sagen, dass Datenschutz den technischen Aspekt des Schutzes der Privatsphäre abdeckt. Juristisch wird Datenschutz, als der Schutz personenbezogener Daten vor missbräuchlicher Verwendung bezeichnet [6]. Um Datenschutz zu quantifizieren, messbar und vergleichbar zu machen und letztendlich dann um Fragen wie "Wie gut hält System XY Datenschutz ein?" beantworten zu können, reicht diese Definition allerdings nicht aus. Um das trotzdem zu können, werden jeweils einzelne Aspekte von *privacy* betrachtet und Metriken darauf angewendet. Metriken bilden mit Hilfe mathematischer Funktionen, Eigenschaften eines Systems auf einen Zahlenwert ab. In dieser Arbeit werden bestehende Metriken vorgestellt, die genau das möglichen machen. Dabei wird zwischen Metriken unterschieden, die sich allgemein anwenden lassen und Metriken, die nur für bestimmten Anwendungen funktionieren. Zunächst wird die rechtliche Situation anhand des Bundesdatenschutzgesetzes in Kapitel 2 erläutert. Besonderes soll gezeigt werden welche Vorgaben zu Datenschutz das Gesetz vorgibt, um dann vergleichen zu können, inwiefern die vorgestellten Metriken dazu beitragen diese Vorgaben zu erfüllen. In Kapitel 3.1 wird eine Metrik vorgestellt, die misst wie gut Informationen über Standorte in *vehicle-to-infrastructure (V2X)* Systemen geschützt werden. In Kapitel 3.2 werden drei verschieden Metriken gezeigt, die dazu dienen den Grad von Anonymität von zu veröffentlichenden Mikrodaten, also Daten, die zu statistischen Zwecken über Individuen erhoben wurden, zu bestimmen. Dabei liegt der Fokus, darauf welche Angriffe die Datensätze, die mit Hilfe der Metriken anonymisiert wurden, verhindern und welche potentielle Probleme nach der Anonymisierung bestehen. In Kapitel 3.2.4 wird die Definition der zuvor vorgestellten Metriken um einen informationstheoretischen Ansatz erweitert und ein Zusammenhang zwischen den Metriken in Kapitel 3.2 hergestellt. In Kapitel 4 geht es dann darum Metriken zu finden, die möglichst allgemein anwendbar sind.

2. RECHTLICHE SITUATION

Der Datenschutz ist in Deutschland seit 1977 gesetzlich im Bundesdatenschutzgesetz (BDSG) [12] geregelt. Unabhängig von der Anwendung schreibt das BDSG in § 3a Datenvermeidung und Datensparsamkeit vor. Dies bedeutet, dass nur so wenig personenbezogenen Daten wie es für die Anwendung nötig ist, zu erheben sind. Ist es trotzdem nötig personenbezogene Daten zu nutzen, so sind diese zu anonymisieren. Dies wird auch nochmals explizit in § 30, §30a und § 40, die die "[g]eschäftsmäßige Datenerhebung und -speicherung zum Zweck der Übermittlung in anonymisierter Form" , "Markt- oder Meinungsforschung" und "Verarbeitung und Nutzung personenbezogener Daten durch Forschungseinrichtungen"

doi: 10.2313/NET-2015-09-1_09

regeln erwähnt. In allen drei Paragraphen wird vom Gesetz verlangt, dass die Datenbestände anonymisiert werden, sobald es "nach dem Zweck des Forschungsvorhabens bzw. nach dem Forschungszweck möglich ist". Der Begriff *Anonymisierung* wird in § 3 (6) definiert:

> 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.

Interessant hierbei ist, dass die einzige Aussage, wie weit die Daten anonymisiert werden müssen, ist, dass sie nur mit unverhältnismäßig großem Aufwand Personen zugeordnet werden können. Das kann unter Umständen relativ frei interpretiert werden. Es liegt also in der Hand der Gerichte, basierend auf aktuelle technische Gegebenheiten, zu entscheiden ob die Anonymisierung ausreichend ist. In den Kommentaren und Erläuterungen zu § 3 Absatz 6 wird allerdings näher erläutert wie die Anonymisierung zu erfolgen hat [13]. Hierbei sind durchaus Parallelen zu den Methoden erkennbar, die im folgenden Kapitel vorgestellt werden. So wird verlangt, dass *identifier* gelöscht werden. Es werden auch explizit *identity disclosure* und *attribute disclosure* erwähnt und dass diese durch verallgemeinerte *quasi identifiers* verhindert werden können. *Quasi identifier* sind Attribute die in Kombination ein Individuum identifizieren können. Auf die Begrifflichkeiten wird in Kapitel 3.2 näher eingegangen. Die genauen Hintergründe hierzu werden ebenfalls in diesem Kapitel erläutert. Zusätzlich zu den dort vorgestellten Methoden wird in dem Gesetzkommentar auch empfohlen Zufallsfehler, quasi ein Rauschen, den Daten hinzuzufügen. Aber auch hier fehlen Angaben, wie wie weit die Verallgemeinerung gehen muss.

3. ANWENDUNGSPEZIFISCHE METRIKEN

3.1 Location privacy metric in V2X

Vehicle-to-infrastructure (V2X) bezeichnet man ein Kommunikationssystem, bei dem Fahrzeuge Informationen mit verkehrstechnischer Infrastruktur austauschen. Je nach Anwendung kann es sein, dass der aktuelle Aufenthaltsort, der als besonders schützenswert gilt, beispielsweise für Funktionen, wie der Kollisionsvermeidung, ziemlich detailliert preisgegeben werden muss. Dieses preisgeben des Standortes kann als Verletzung der Privatsphäre gesehen werden. Wie detailliert die Informationen über den Standort einer Person preisgegeben werden, kann in dieser Metrik festgehalten werden. Dazu sind einige Modellierungen nötig. Wir betrachten Fahrten, die aus einer Serie von Aufenthaltsorten bestehen. Gemessen wird die Möglichkeit eines Angreifers eine Fahrt einer bestimmten Person zuzuordnen. Zum einen modellieren wir einen Graphen, der drei verschiedene Arten von Knoten besitzt. Individuen I, Ursprung O und Ziel D. Wir nehmen an, dass es gleich viele Ursprünge und Ziele gibt. Die Kanten sind mit einer Wahrscheinlichkeit p gewichtet. Zusätzlich zu dem Graphen lässt sich das Modell auch mit drei Adjazenzmatrizen IO, OD und DI repräsentieren. Die Einträge in IO stellen die Wahrscheinlichkeiten dar ein Individuum einem Ursprung zuzuordnen, in OD sind die Einträge die Wahrscheinlichkeiten für Fahrten zwischen O und D. In DI sind entsprechend die Wahrscheinlichkeiten ein Individuum einem Ziel zuzuordnen. Die Summen der Zeilen in OD und DI müssen genau 1 ergeben. In IO darf die Summe einer Zeile auch kleiner 1 sein. Die verbleibende Differenz zu 1 ist die Wahrscheinlichkeit, dass das in der entsprechenden Zeile repräsentierte Individuum keine Fahrt antritt und daheim bleibt. Diese Wahrscheinlichkeit wird als p^c bezeichnet. Abbildung 1 (a) zeigt beispielhaft den Graph der auf ein Individuum zentriert wurde. In Abbildung (b) wurde der Graph dahingehend vereinfacht, dass nur noch das Produkt aus den Einzelwahrscheinlichkeiten angezeigt wird.

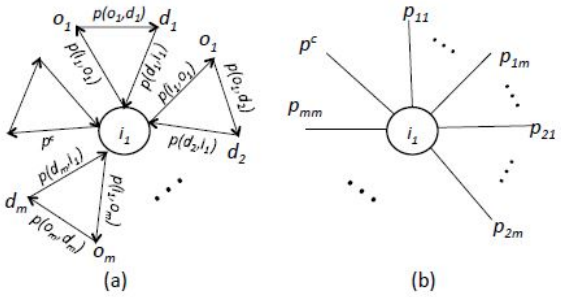

Abbildung 1: Beispielgraph [7]

Um die Informationen die in den Wahrscheinlichkeiten stecken messbar zu machen, bedient man sich der Entropie. Entropie ist ein Maß um anzugeben wie viel Unsicherheit eine Verteilung enthält. Je höher die Unsicherheit ist, desto höher ist die Entropie. Eine hohe Entropie ist im Sinne des Datenschutzes erstrebenswert. Die Entropie H ist folgendermaßen definiert:

$$H = -\sum p_i log_b p_i \tag{1}$$

Wird der Logarithmus zur Basis b=2 genommen, ist die Einheit der Entropie *bit*. Die Entropie $H(i_s)$ für ein Individuum i_s wird wie folgt definiert:

$$H(i_s) = -(\sum_{j=1}^{m}\sum_{k=1}^{m} \hat{p}_{jk} log(\hat{p}_{jk}) + \hat{p}^c log(\hat{p}^c)) \tag{2}$$

Dabei sind $\hat{p}_{jk}$ und $\hat{p}^c$ die normalisierten Wahrscheinlichkeiten, dass i_s eine Fahrt macht bzw. keine Fahrt macht.

$$\hat{p}_{jk} = \frac{p(i_s, o_j)p(o_j, d_k)p(d_k, i_s)}{\sum_{j=1}^{m}\sum_{k=1}^{m} p(i_s, o_j)p(o_j, d_k)p(d_k, i_s) + \hat{p}^c} \tag{3}$$

$$\hat{p}^c = 1 - \sum_{j=1}^{m} p(i_s, o_j) \tag{4}$$

Um den Wert der hier berechnet wird besser einordnen zu können, wird zusätzlich die maximal mögliche Entropie $MaxH$ berechnet und in Relation zu der für ein Individuum berechneten Entropie gesetzt:

$$H_{\%} = \frac{H(i_s)}{MaxH(i_s)} 100\% \tag{5}$$

$$MaxH(i_s) = -log(\frac{1}{m^2+1}) \tag{6}$$

doi: 10.2313/NET-2015-09-1_09

$H_{\%}$ gibt somit an wie weit der Datenschutz für ein Individuum vom maximal möglichen Datenschutz entfernt ist. Diese Definition besitzt einige Parallelen zu den Definition von *unlinkability* bzw. *anonymity*, die im Kapitel 4 vorgestellt werden. Dies liegt allerdings weniger daran, dass man diese Metrik auf einen allgemeinen Fall zurückführen kann, sondern dass die Metrik in diesem Kapitel auf diesem Prinzip aufbaut. Man kann eher davon sprechen, dass die *location privacy metric* ein Spezialfall der *unlinkability* Metrik ist, der entsprechend den Eigenheiten, die $V2X$ mit sich bringt, angepasst wurde.

3.2 Metriken zur Anonymisierung von Mikrodaten

Der Anwendungsfall, dass Datensätze anonym veröffentlicht werden sollen, die sensible Daten enthalten, ist der vermutlich am besten erforschte Bereich, was *privacy*-Metriken angeht. So gibt es in der Konsequenz bereits einige verschiedene Metriken, die ausdrücken wie stark die Daten anonymisiert sind und die Anhaltspunkte geben, ob die Daten weiter anonymisiert werden müssen, um die Privatsphäre der einzelnen Personen zu gewährleisten und wie hoch das Risiko einer Deanonymisierung ist.

Nach [10] können die Daten aus den zu veröffentlichenden Datensätzen in drei Kategorien eingeteilt werden. Zum einen gibt es Attribute, die die Individuen eindeutig identifizieren können. Beispielsweise die Passnummer, die Sozialversicherungsnummer oder auch der vollständige Name. Diese Attribute werden *Identifiers* genannt. Attribute die in Kombination, aber nicht alleine, ein Individuum identifizieren können werden *quasi-identifiers (QI)* genannt. Das können z.B. Alter, Geschlecht, usw... sein. Attribute die sensitive Daten über Individuen, z.B. Krankheitsbefunde, enthalten, werden *sensitive attributes* genannt. Ein Datensatz kann mehrere verschiedene *sensitive attribute* enthalten. Der Einfachheit halber, gehen wir in dieser Arbeit davon aus, dass der Datensatz nur eines enthält. Betrachten wir nun als Beispiel *Tabelle 1*, welches aus [10] entnommen ist und für unsere Bedürfnisse angepasst wurde. Dann ist dort die Diagnose W das *sensitive attribute* und Namen und Höhe X sind jeweils *QI*. Dieser Datensatz soll nun veröffentlicht werden. Dazu muss der Datensatz anonymisiert werden, um die Privatsphäre der Patienten zu schützen. Hierfür werden in Kapitel 3.2.1, Kapitel 3.2.2 und Kapitel 3.2.3 jeweils verschiedene Metriken herangezogen.

3.2.1 k-anonymity

Die k-anonymity Metrik ist die am wenigsten strenge Metrik. Ziel dieser Metrik ist es so genannte *linking attacks*, bei denen der Angreifer versucht durch die Kombination der veröffentlichten *quasi-identifiers* Rückschlüsse auf die Identität der Personen zu erlangen. Die Metrik besagt, dass für jede Kombination von *quasi-identifier* es mindestens k Einträge in der Tabelle gibt. Jeder Eintrag ist dann einer Person mit einer Wahrscheinlichkeit von $1/k$ zuordenbar. In Bezug auf Tabelle 1 würde diese Vorgabe, der k-Anonymität, bedeuten, dass es für jede Kombination aus Namen und Größe X mindestens k Einträge geben muss. Um dies zu erreichen wird in der anonymisierten Tabelle mit $k = 2$ der Name komplett zensiert und statt der Größe X werden nur noch Bereiche die jeweils 10cm umfassen angegeben. Die Daten wurden also

Original Dataset {*X, W*}		
Name	Height X	Diagnosis W
Timothy	166	N
Alice	163	N
Perry	161	N
Tom	167	N
Ron	175	N
Omer	170	Y
Bob	170	N
Amber	171	N
Sonya	181	N
Leslie	183	N
Erin	195	Y
John	191	Y

Tabelle 1: Original Datensatz [10]

generalisiert. Tabelleneinträge, die dieselben Werte für die *quasi-identifier* besitzen, gehören einer Äquivalenzklasse an. Die Äquivalenzklassen werden auch einfach Blöcke genannt. Wäre der Datensatz größer, hätte man eventuell auch den Anfangsbuchstaben oder noch mehr von den Namen nicht zensieren müssen oder auch die Bereiche der Größe kleiner fassen können. Der Umfang der Anonymisierungen hängt also nicht nur vom gewählten k ab, sondern auch von der Menge der Daten im Datensatz, sowie ihrer konkreten Ausprägung.

Anonymized Dataset {X, W}		
Name	Height $\tilde{X}$	Diagnosis W
******	[160-170]	N
******		N
******		N
******		N
******	[170-180]	N
******		Y
******		N
******		N
******	[180-190]	N
******		N
******	[190-200]	Y
******		Y

Tabelle 2: Anonymisierter Datensatz [10]

Man erkennt schnell, dass der Informationsgehalt mit steigendem k verloren geht. So kann es sein, dass unter Umständen die Daten in ihrem ursprünglichen Verwendungszweck gar nicht mehr verwertbar sind.

Grundsätzlich gibt es zwei verschiedene Risiken der Identifizierung, zum einen *identity disclosure*, welches auftritt, wenn ein Eintrag in der Tabelle einer bestimmten Person zugeordnet werden kann. Dieses Risiko kann mit *k-anonymity* erfolgreich ausgeschlossen werden. Für das zweite Risiko, der *attribute disclosure*, bei dem Attribute einer Person zugeordnet werden können, bedarf es weitere Metriken. Diese werden in den folgenden Kapiteln erläutert. [8]

3.2.2 l-diversity

Bei Betrachtung des letzten Blocks von Tabelle 2 fällt auf, dass das *sensitive attribute* bei allen Einträgen gleich ist.

doi: 10.2313/NET-2015-09-1_09

Dies ermöglicht *homogeneity attacks* und *background attacks*. Der Angreifer kann durch Hintergrundwissen das *sensitive attribute* einer bestimmten Person eindeutig bestimmen. In unserem Beispiel genügt es zu wissen, dass eine Person in dem veröffentlichten Datensatz erfasst ist und dass die Person größer als 190 cm ist, um zur der Erkenntnis zu gelangen, dass diese Person über einen positiven Befund verfügt. In [9] wird eine neues Prinzip, *l-diversity*, eingeführt, um dieser Art von Angriffen entgegenzuwirken. Es gibt verschiedene Möglichkeiten *l-diversity* zu definieren. grundsätzlich schreibt diese Metrik vor, dass in jedem Block die l häufigsten Werte der *sensitive attribute*, mindestens einmal, idealerweise jedoch möglichst gleich verteilt, repräsentiert werden. Ein Block erfüllt *l-diversity*, wenn mindestens l Werte vorkommen. Ein Datensatz erfüllt *l-diversity*, wenn alle Blöcke *l-diversity* erfüllen. Unser Beispiel kann maximal *2-diversity* erfüllen, da das *sensitive attribute* nur 2 Werte annehmen kann. Betrachtet man die Tabelle 2, stellt man fest, dass nicht mal *2-diversity* erfüllt ist. Nur im 2. Block kann man von *2-diversity* sprechen.

3.2.3 t-closeness

Dass *l-diversity* nicht immer zielführend ist, kann man ebenfalls wieder in unserem Beispiel erkennen. In dem Beispiel kann das *sensitive attribute* nur zwei Werte annehmen. Je nach Befund, positiv oder negativ. Zum anderen kommt der positive Befund gewöhnlich viel seltener vor, als der negative. So auch in diesem Beispiel. In Blöcke in denen kein positiver Wert vorliegt, führt das zu keinen Problemen, da man davon ausgehen kann, mit einem negativen Befund in Verbindung gebracht zu werden nicht die Privatsphäre eines Einzelne verletzt. Es kann allerdings per Ausschlussprinzip zu Verletzung der Privatsphäre eines dritten kommen. Angenommen in einem Block befinden sich gleich viele positive und negative Befunde, dann erfüllt dieser zwar *2-diversity*, aber jede Person, deren Daten in diesem Block repräsentiert werden, wird plötzlich mit einer Wahrscheinlichkeit von $p = 50\%$ mit einem positiven Befund in Verbindung gebracht, obwohl die Wahrscheinlichkeit eines solchen Befundes sehr viel geringer ist. Dieses Problem ermöglicht sogenannte *skewness attacks*. Wenn das *sensitive attribute*, anderes als in unserem Beispiel, mehrere Werte annehmen kann, so kann dies zu *similarity attacks* führen. *l-diversity* berücksichtigt nämlich nicht, ob *sensitive attributes* sich semantisch ähnlich sind. In [11] wird eine dritte Metrik, *t-closeness*, beschrieben, die sich diesen Problemen widmet. Diese Metrik basiert bereits auf einem Prinzip, auf dem auch allgemeinere Metriken basieren, wie sie in Kapitel 4 beschrieben werden. Die a posteriori Wahrscheinlichkeit ein *sensitive attribute* einer Person zuzuordnen, nachdem der Datensatz veröffentlicht wurde, darf nicht größer sein, als die a priori Wahrscheinlichkeit vor der Veröffentlichung. Das wird laut *t-closeness* erreicht, wenn sich die Verteilung des *sensitive attribute* in einem Block maximal durch einen Schwellwert t von der Verteilung des ganzen Datensatzes unterscheidet. Um diesen Unterschied in den Verteilungen zu messen hat sich die *Earth Mover's distance* als am brauchbarsten herausgestellt [11]. Diese misst den Aufwand der nötig wäre, eine Verteilung in die andere zu überführen. Bei der Festlegung von t gilt es abzuwägen, wie viel Informationsverlust man in Kauf nimmt. Denn der Grad an Datenschutz den einem t zusichert steht im direkten Zusammenhang mit dem Grad an Informationen, die verloren gehen. Das lässt sich damit begründen, dass die Informationen der veröffentlichten Datensätze genau in den Unterschieden der Verteilungen liegen. In dem Beispiel könnten Forscher beispielsweise versuchen einen Zusammenhang zwischen der Größe X und der Diagnose W herzustellen. Dieser Zusammenhang geht allerdings mit zunehmender Anonymisierung verloren.

3.2.4 One-symbol information

Da die in diesem Kapitel gezeigten Metriken auf den Schutz vor verschieden Arten von Angriffen abzielen sind sie untereinander nur schwer vergleichbar. Deshalb werden in [10] die Metriken aufgegriffen und auf eine *one-symbol information* Einheit zurückgeführt. Dazu werden *k-anonymity*, *l-diversity* und *t-closeness* mit Hilfe von Entropie und Transinformation neu definiert. Transinformation $I(X;Y)$ gibt die Stärke des statistischen Zusammenhangs zwischen den Zufallsvariablen X und Y an.

$$\begin{aligned} I(X;Y) = & \sum_{x\in X, y\in Y} p(x,y) log_2 \left[\frac{p(x,y)}{p(x)p(y)}\right] \\ = & \sum_{x\in X, y\in Y} p(y)p(x|y) log_2 \left[\frac{p(x|y)}{p(x)}\right] \\ = & H(Y) - H(Y|X) = \sum_{x\in X} p(x)\left[H(Y) - H(Y|x)\right] \end{aligned} \tag{7}$$

Bei *one-symbol information* wird zunächst nicht die ganze Tabelle betrachtet, sondern jeder Eintrag einzeln. Sei x ein Eintrag in einem Datensatz X und $\tilde{x}$ ein Eintrag in einem anonymisierten Datensatz $\tilde{X}$. Die Wahrscheinlichkeit einen Eintrag x mit einem gegebenem $\tilde{x}$ zu identifizieren beträgt $p(x|\tilde{x}) = 1/N_{\tilde{x}}$, wobei $N_{\tilde{x}}$ die Anzahl der Eintrage x ist, die das gegebene $\tilde{x}$ in X abdeckt. N ist dagegen die Anzahl der unterscheidbaren Einträge in X.

k-anonymity

Somit lässt sich *k-anonymity* folgendermaßen mit Hilfe der bedingten Entropie ausdrücken:

$$H(X|\tilde{x}) \geq log_2 k \tag{8}$$

Für *One-symbol information* gibt es vier verschiedene Definitionen I_1, I_2, I_3 und I_4, wobei hier nur die ersten zwei Definitionen verwendet werden. Als *one-symbol information* wird *k-anonymity* dann so definiert:

$$I_2(X,\tilde{x}) \equiv H(X) - H(X|\tilde{x}) \leq log_2 \frac{N}{k} \tag{9}$$

Will man nicht die einzelnen Einträge $\tilde{x}$ sondern den ganzen Datensatz $\tilde{X}$ betrachten muss man den Durchschnitt betrachten, der wie folgt definiert wird.

$$I(X,\tilde{X}) \leq log_2 \frac{N}{k} \tag{10}$$

l-diversity

l-diversity kann ebenfalls mittels Entropie ausgedrückt werden:

$$H(W|\tilde{x}) \geq log_2 l \tag{11}$$

W ist dabei das *sensitive attribute*. Als *one-symbol information* ausgedrückt lautet die Formel:

$$I_2(W,\tilde{x}) \equiv H(W) - H(W|\tilde{x}) \leq H(W) - log_2 l \tag{12}$$

doi: 10.2313/NET-2015-09-1_09

Als Durchschnitt über $\tilde{X}$:

$$I(W, \tilde{X}) \equiv H(W) - H(W|\tilde{X}) \leq H(W) - log_2 l \quad (13)$$

t-closeness

Auch *t-closeness* lässt sich als *one-symbol information* ausdrücken:

$$I_1(W, \tilde{x}) \equiv \sum_{w \in W} p(w|\tilde{x}) log_2 \frac{p(w|\tilde{x})}{p(w)} \leq t \quad (14)$$

Als Durchschnitt über $\tilde{X}$:

$$I_1(W, \tilde{X}) \equiv \sum_{\tilde{x} \in \tilde{X}} p(\tilde{x}) \sum_{w \in W} p(w|\tilde{x}) log_2 \frac{p(w|\tilde{x})}{p(w)} \leq t \quad (15)$$

Aus diesen Definition lassen sich nun die oberen und unteren Schranken für l herleiten, als auch ein Zusammenhang zwischen l und t herstellen:

$$1 \leq l \leq l_{max} \equiv 2^{H(W)} \quad (16)$$

$$l_t = 2^{H(W)-t} \quad (17)$$

So lässt sich mit l_t ein l für ein gegebenes t berechnen.

4. ALLGEMEINE METRIKEN

Da sich die Metriken aus Kapitel 3 nur jeweils für einen speziellen Anwendungsfall nutzen lassen, wird in diesem Kapitel versucht Metriken aufzustellen, die sich unabhängig davon, welcher Anwendungsfall vorliegt, immer anwenden lassen sollen. In [14] werden zwei Metriken, *degree of anonymity* und darauf basierend *degree of unlinkability* eingeführt, die genau das zum Ziel haben. In diesem Kapitel wird deshalb näher auf diese Metriken eingegangen.

4.1 Degree of anonymity

In 3.2.3 hat man a priori und a posteriori Wahrscheinlichkeiten verglichen, um zu sehen an wie viel Informationen ein Angreifer gelangen kann. Auf einem ähnlichen Prinzip basiert der *degree of anonymity*. Während bei *t-closeness*, wie bereits in Kapitel 3.2.3 erwähnt, die *Earth Mover's Distance* benutzt, welche den Aufwand beschreibt, eine Verteilung in eine andere zu überführen, werden beim *degree of anonymity* die aus der Informationstheorie bekannten Entropien verglichen. Dafür wird zunächst $A = \{a_1, \ldots, a_n\}$ als eine nicht-leere, endliche Menge von Aktionen und $U = \{u_1, \ldots, u_n\}$ als eine Menge von Benutzer definiert. Jedes $u_i \in U$ mit $i \in \{1, \ldots, n\}$ führt a mit einer Wahrscheinlichkeit $p_i > 0$ aus. Die a priori Wahrscheinlichkeit, dass u_i die Aktion a ausgeführt hat, beträgt idealerweise $1/n$. Durch Beobachtung kann der Angreifer Rückschlüsse ziehen und auf eine a posteriori Wahrscheinlichkeit, die sich von der a priori Wahrscheinlichkeit unterscheidet, schließen. Die a posteriori Wahrscheinlichkeit wird mit Hilfe der Zufallsvariable X definiert, wobei $p_i = P_a(X = u_i)$ gilt. Der *degree of anonymity* lässt sich berechnen wenn man den Unterschied zwischen der maximal möglichen Entropie $max(H(X)) = log_2(n)$ und der a posteriori Entropie $H(X) = -\sum_{i=1}^{n} p_i log_2(p_i)$ betrachtet. Weil man nicht die Größe der Menge U, sondern nur die Verteilung messen will, ist es nötig den Unterschied zusätzlich noch zu normalisieren. Der *degree of anonymity* $d(U)$ wird folglich folgendermaßen definiert:

$$d(U) := 1 - \frac{max(H(X)) - H(X)}{max(H(X))} = \frac{H(X)}{max(H(X))} \quad (18)$$

Durch die Normalisierung nimmt $d(U)$ nur Werte im Bereich $[0, 1]$ an. $d(U) = 0$ bedeutet, dass es ein Subjekt gibt, welches einer Aktion mit der Wahrscheinlichkeit $p_i = 1$ zuzuordenbar ist. $d(U) = 1$ dagegen würde bedeuten, dass jede Aktion a einem Subjekt nur mit einer Wahrscheinlichkeit von $1/n$ zuordenbar ist [14].

4.2 Degree of unlinkability

Das Konzept der Anonymität ist nicht immer ausreichend, weil es nur auf Personen anwendbar ist. Deshalb ist es sinnvoll zusätzlich den *Degree of unlinkability* zu definieren. *Unlinkability* wird folgendermaßen definiert: Zwei Elemente sind, nachdem man ein System beobachtet hat, nicht einander mehr oder auch weniger zuordenbar, als zuvor. Unterscheidet sich die a priori Wahrscheinlichkeit, Elemente einander zuordnen zu können, von der a posteriori Wahrscheinlichkeit, spricht man von einem *existential break*. Die Elemente können dabei alles mögliche sein. Zum Beispiel Personen, Nachrichten, Aktionen oder vieles mehr. Für die formale Definition wird wieder eine Menge $A = \{a_1, \ldots, a_l\}$ mit den Elementen aus dem zu betrachteten System definiert. Innerhalb dieser Menge werden die Äquivalenzklassen $A_1, \ldots, A_n$ gebildet, die jeweils Elemente enthalten, die einander verwandt sind, beispielsweise Nachrichten die vom gleichen User versendet wurden. Außerdem wird die Äquivalenzrelation $\sim_{r(A)}$ gebildet, welche bedeutet 'ist verwandt mit'. Zunächst betrachten wir den Fall, dass der *degree of unlinkability* von zwei Elementen a_i und a_j innerhalb einer Menge berechnet wird. Der *degree of unlinkability d(i,j)* zweier Elemente a_i und a_j wird wie folgt definiert:

$$\begin{aligned} d(i,j) := \quad & H(i,j) \\ = \quad & -P(a_i \sim_{r(A)} a_j) \cdot log_2(P(a_i \sim_{r(A)} a_j)) \\ & -P(a_i \not\sim_{r(A)} a_j) \cdot log_2(P(a_i \not\sim_{r(A)} a_j)) \end{aligned} \quad (19)$$

$d(i,j) = 0$ bedeutet, dass der Angreifer zu 100% a_i und a_j derselben Äquivalenzklasse zuordnen kann oder ausschließen kann, dass sie in derselben sind. $d(i,j) = 1$ dagegen bedeutet, dass der Angreifer a_i und a_j mit einer Wahrscheinlichkeit von $\frac{1}{2}$ derselben Äquivalenzklasse zuordnen kann oder eben nicht. Im nächsten Schritt wird die Definition dahingegen erweitert, dass nicht die Zuordenbarkeit von zwei, sondern beliebig vielen Elementen bestimmt werden kann. Sei $\{a_{i_1}, \ldots, a_{i_k}\}$ eine Teilmenge von A mit $2 < k \leq n$ Elementen. $P((\sim_{r_j(A)} |\{a_{i_1}, \ldots, a_{i_k}\}) = (\sim_{r(A)}))$ ist dann die Wahrscheinlichkeit, dass die betrachteten Elemente $\{a_{i_1}, \ldots, a_{i_k}\}$ den richtigen Äquivalenzklassen aus A zugeordnet wurden. Die Definition des *degree of unlinkability* $d(i_1, \ldots, i_k)$ lautet dann folglich:

$$\begin{aligned} d(i_1, \ldots, i_k) := \quad & H(i_1, \ldots, i_k) \\ = \quad & -\sum_{j \in I_k} \frac{1}{|I_k|} \\ & \cdot[P((\sim_{r_j(A)} |\{a_{i_1}, \ldots, a_{i_k}\}) = (\sim_{r(A)})) \\ & \cdot log_2(P((\sim_{r_j(A)} |\{a_{i_1}, \ldots, a_{i_k}\}) = (\sim_{r(A)})))] \end{aligned} \quad (20)$$

doi: 10.2313/NET-2015-09-1_09

Dabei ist I_k ein Index, der alle möglichen Äquivalenzklassen zählt und $|I_k| = 2^{k-1}$.

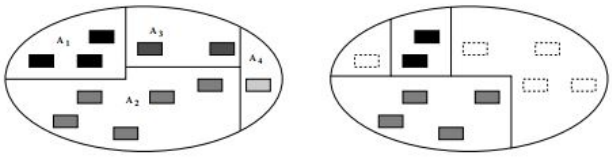

Abbildung 2: Beispiel: Teilmenge [14]

Abbildung 2 ist ein Beispiel für die zuvor beschriebene Situation. Links befindet sich die Menge A, die sich in vier Äquivalenzklassen, A_1 bis A_4, aufteilen lässt. Rechts wurde eine Teilmenge von A den Äquivalenzklassen zugeordnet.

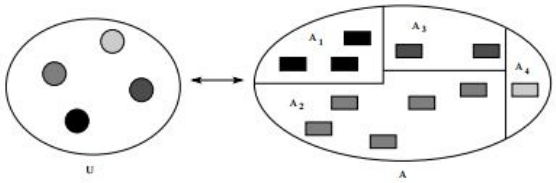

Abbildung 3: Beispiel: Mehrere Mengen [14]

Je nach Szenario, wie z.B. in Kapitel 4.1, kann es nötig sein mehr als eine Menge zu modellieren. Beispielsweise können die Benutzer in einer Menge U zusammengefasst werden und die Aktionen in einer anderen Menge A zusammengefasst werden. Abbildung 3 illustriert dieses Szenario. Die Definitionen lassen sich dahingehend problemlos anpassen, um auch diesen Fall abzudecken [14].

Die zwei Metriken lassen sich grundsätzlich bei allen Systemen anwenden. Die Herausforderung besteht dabei die zu untersuchenden Systeme korrekt zu modellieren. Da die Metriken sehr allgemein gehalten sind, können die Ergebnisse unter Umständen nicht zielführend genug sein. In diesem Fall bieten sich die Metriken aber an, um auf ihnen basierend eine spezifischere Metrik zu definieren, wie es beispielsweise auch in Kapitel 3.1 getan wurde.

5. FAZIT

Anhand von verschiedenen Metriken konnte gezeigt werden, dass gewisse Aspekte von *privacy* durch Metriken durchaus messbar sind. Die meisten Metriken zeigen auf, ob der Datenschutz in den untersuchten Systemen gewährleistet wird. Der Großteil bietet allerdings keinen Lösungsansatz, wie die Situation zu verbessern wäre. Sie dienen deshalb hauptsächlich als Indikatoren, ob es zu Problemen mit Datenschutz geben kann. Nur die in Kapitel 3.2 vorgestellten Metriken können auch dazu verwendet werden den Datenschutz, in diesem Fall speziell die Anonymität, zu verbessern. Viele der verfügbaren Metriken sind für konkrete Anwendungsfälle konzipiert. Deshalb sind die Ergebnisse aus den Metriken nur eingeschränkt untereinander vergleichbar. Es konnte aber auch gezeigt werden, dass es Metriken gibt, die grundsätzlich für alle Anwendungen anwendbar sind. Es wäre allerdings wünschenswert, dass es in diesem Bereich weitere Metriken gibt, die größere Teile von *privacy* abdecken. In den bestehenden Metriken wurde hauptsächlich der Aspekt der Anonymität und *unlinkability* beleuchtet. Gerade wenn man das Bundesdatenschutzgesetz [12] betrachtet, fällt auf, dass dort bei der Verarbeitung und Veröffentlichung der persönlichen Daten die Zustimmung hierzu eine große Rolle spielt. Dies wurde in den gezeigten Metriken gar nicht berücksichtigt. Deshalb wäre es abschließend auch wünschenswert, dass Metriken entwickelt werden, die diesen Aspekt mit aufnehmen.

6. LITERATUR

[1] *dict.cc, privacy*, `https://www.dict.cc/?s=privacy`, (Abgerufen am 17. 07. 2015)

[2] *leo.org, privacy*, `http://dict.leo.org/ende/index_en.html#/search=privacy`, (Abgerufen am 17. 07. 2015)

[3] *Duden, Datenschutz*, `http://www.duden.de/rechtschreibung/Datenschutz`, (Abgerufen am 17. 07. 2015)

[4] *Allgemeine Erklärung der Menschenrechte*, `http://www.un.org/depts/german/menschenrechte/aemr.pdf`, (Abgerufen am 17. 07. 2015)

[5] *Grundgesetz*, `http://www.gesetze-im-internet.de/gg/BJNR000010949.html#BJNR000010949BJNG0001` (Abgerufen am 17. 07. 2015)

[6] *Datenschutz: Definition, Begriff und Erklärung*, `http://www.juraforum.de/lexikon/datenschutz`, (Abgerufen am 17. 07. 2015)

[7] Ma, Z., Kargl, F. and Weber, M.: *A location privacy metric for V2X communication systems* (2009)

[8] Samarati, P. and Sweeney, L.: *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.*, Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory (1998)

[9] Machanavajjhala A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. : *l-Diversity: Privacy Beyond k-Anonymity*, ACM Trans. Knowl. Discov. Data 1, 1, Article 3 (2007)

[10] Bezzi, Michele: *An information theoretic approach for privacy metrics*, TRANSACTIONS ON DATA PRIVACY 3 (2010)

[11] Li, N., Li, T. and Venkatasubramanian, S.: *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity*, ICDE. Vol. 7. (2007)

[12] *Bundesdatenschutzgesetz*, `http://http://www.gesetze-im-internet.de/bdsg_1990/`, (Abgerufen am 17. 07. 2015)

[13] *Kommentare und Erläuterungen zu § 3 Weitere Begriffsbestimmungen*, `http://www.bfdi.bund.de/bfdi_wiki/index.php/3_BDSG_Kommentar_Absatz_6`, (Abgerufen am 17. 07. 2015)

[14] Steinbrecher S. and Köpsell, S.: *Modelling Unlinkability* (2003)

doi: 10.2313/NET-2015-09-1_09