Allgemeine Konzepte
K-Anonymity, l-Diversity and T-Closeness

Dietmar Hauf

IPD Uni-Karlsruhe

Zusammenfassung Die Veröffentlichung von personenbezogenen Daten über eine Menge an Individuen unter Einhaltung der Privatsphäre des Einzelnen ist ein wichtiges Problem im Informationszeitalter. Diese Ausarbeitung weist auf die Problematik mit zu verarbeitenden Daten und dem Schutz der Privatsphäre enthaltener Individuen hin. Hierbei werden drei unterschiedliche Sicherheitsmaße vorgestellt. Zunächst wird auf k-Anonymity eingegangen. Diese verhindert die Reidentifizierung einer Person in einem Datenbestand durch Anonymisierung dieser in einer Gruppe von anderen Personen des Datenbestandes. Es werden zwei Angriffsszenarien auf k-Anonymity vorgestellt, welche jedoch mit einfachen Mitteln umgangen werden können. Wie zwei weitere Angriffsszenarien zeigen werden, stellt k-Anonymity als Schutzmaß der Privatsphäre nicht zufrieden. Anschließend wird l-Diversity vorgestellt, welches eben diese Lücken schließt. Hierbei wird k-Anonymity um ein Maß dem Schutz vor Identifizierung spezieller Attribute des Individuums erweitert. Aber auch dies ist kein ausreichendes Qualitätsmerkmal, um alle Aspekte der Privatsphäre zu schützen, wie es weitere Angriffe verdeutlichen werden. Eine Lösung und gleichzeitig ein Ersatz für l-Diversity bietet das dritte Konzept t-Closeness. Nach der Vorstellung dieser Konzepte wird auf den Aspekt der Weiterverwendung der Daten eingegangen. Es wird gezeigt, dass hierbei der Schutz der Privatsphäre mit Informationsverlust einhergeht. Schließlich wird eine Lösung, die beide Aspekte berücksichtigt, veranschaulicht.

Inhaltsverzeichnis

Allgemeine Konzepte 1
Dietmar Hauf
1 Einführung 2
2 Die Problematik in Bezug auf den Einzelnen 3
3 K-Anonymity 5
3.1 Angriffe gegen k-Anonymity 6
3.2 Grenzen von k-Anonymity 7
4 L-Diversity 8
4.1 Prinzip der l-Diversity 8
4.2 konkrete l-Diversity Instanzen 9
4.3 Grenzen der l-Diversity 11
5 T-Closeness 12
6 Probleme der vorgestellten Konzepte 13
6.1 Anatomy 15
7 Fazit 16

1 Einführung

Wir leben heutzutage im Informationszeitalter - von E-mailkommunikation bis zu intelligenten Kühlschränken - die elektronische Datenverarbeitung ist längst Alltag geworden. Viele Organisationen wie Krankenhäuser, Kreditunternehmen oder Einkaufszentren erfassen und verwalten personenbezogene Daten. Kauft man beispielsweise eine teure antike Vase mit seiner Visa-Karte ein, sollte man nicht verwundert sein, wenn man innerhalb weniger Minuten einen Anruf von seiner Bank erhält, vorausgesetzt diese Karte wurde seither nur für alltägliche Essenseinkäufe benutzt. So würden bei Visa sofort „Alarmlämpchen“ blinken, da dieser Einkauf nicht seinem Kundenverhalten entspricht und sie sich den Einkauf zur Sicherheit von einem nochmals bestätigt lassen möchten. Im Prinzip eine positive Eigenschaft nur sollte man sich spätestens da Gedanken über die eigene Privatsphäre machen. - Wie sicher sind diese Daten? - Möglicherweise schickt die Bank diese Daten an eine Dataminingfirma zur Analyse kreditwürdiger Kunden. Aber nicht nur hier können personenbezogene Daten verbreitet werden. Das statistische Bundesamt Deutschland sammelt beispielsweise alle gesundheitlichen, finanziellen oder andere statistisch relevante Daten über die Bevölkerung Deutschlands. Diese werden dann sogar der Öffentlichkeit zugänglich gemacht (vgl. http://www.destatis.de/jetspeed/portal/cms/). Wie kann also der Schutz der Privatsphäre gewährleistet werden? Es wird somit klar, dass wir eine einzuhaltende Kenngröße für den individuellen Schutz in einer Ansammlung personenbezogener Daten benötigen. Anhand von einer Beispieltabelle (1) mit Daten eines Krankenhauses, die dem statistischen Bundesamt Deutschlands übermittelt werden soll, werden im Zuge dieser Ausarbeitung unterschiedliche

Konzepte als Qualitätsmaß zur Sicherstellung der Privatsphäre veranschaulicht. Aus Sicht des Individuums sollte dabei dem ausreichenden Schutz seiner personenbezogenen Daten die größte Aufmerksamkeit zukommen. Dieses steht jedoch in Diskrepanz mit dem Wunsch nach Information für statistische Weiterverarbeitung der Daten. Darauf wird in Kap.6 näher eingegangen.

Name	Geburtstag	Geschlecht	PLZ	Krankheit
Meisler Hans	17.04.75	M	76227	Impotenz
Witzig Peter	31.07.75	M	76228	Hodenkrebs
Feldlager Karl	17.01.75	M	76227	Sterilität
Schwäger Till	05.07.83	M	76133	Schizophrenie
Traube Kai	31.12.81	M	76139	Diabetes
Kneighty Kaira	05.07.83	W	76133	Magersucht
Krempel Naomi	31.10.83	W	76131	Magersucht

Tabelle 1. medizinische Tabelle

2 Die Problematik in Bezug auf den Einzelnen

Ziel dieser Ausarbeitung ist es, unterschiedliche Qualitätsmerkmale, die den Schutz der personenbezogenen Daten von beispielsweise einer medizinische Tabelle (1) bewerten, aufzuzeigen.
Wie zu erkennen ist, bietet diese Form der medizinischen Tabelle (1) noch keinen Schutz der Privatsphäre. Solange personenspezifische Attribute wie Name und Nachname darin enthalten sind, kann kein Schutz gewährleistet werden. Unter personenspezifischen Attributen versteht man Attribute, die einem Individuum eindeutig zugeordnet sind. Darunter fallen neben dem vollständigen Namen auch die vollständige Adresse und sämtliche staatlichen Kennnummern wie z.B. Sozialversicherungsnummer.
Ein gängiges Verfahren anonymisiert also zunächst die Daten, in dem es zum Schutz der Individuen alle personenspezifischen Attribute aus einer Tabelle entfernt. So wird die Tabelle T (2) durch Anonymisierung der medizinischen Tabelle (1) gewonnen. Die resultierenden Daten scheinen einem Betrachter auf den ersten Blick keinen direkten Bezug zu Personen mehr herzustellen, was jedoch nicht der Fall ist. Einige wenige Charakteristiken einer Person können ausreichen, sie eindeutig aus der Masse zu identifizieren. Ist dies der Fall, kann so ein Tuppel einer Person zugeordnet und die Privatsphäre betreffenden Attribute ausgelesen werden. Die Tragweite dieses Problems verdeutlicht die Publikation [2]. In dieser Veröffentlichung wird unter anderem anhand einer Studie darauf hingewiesen, dass 87% der amerikanischen Bevölkerung eindeutig anhand von nur drei nicht personenspezifischen Attributen identifiziert werden können. Attributwertkombination von Geburtsdatum, ZIPCode und Geschlecht können somit 216 Millio-

nen Menschen aus einer Menge von 248 Millionen eindeutig zugeordnet werden. Solche Sets an Attributen werden in [10] als Quasi-Identifier benannt welcher wie folgt definiert ist:

Definition 1. *Gegeben sei eine Population aus Individuen U*
Eine personenbezogene Tabelle $T\,(A_1, ..., A_n)$
Die endliche Menge an Attributen von $T\,\{A_1, ..., A_n\}$
Funktionen $f_c : U \to T$ *und* $f_g : T \to U^{'}$ $U^{'} \subseteq U$.
Dann ist ein Quasi-Identifier von T, geschrieben Q_T*, ein Set aus Attributen* $\{A_i, ..., A_j\} \subseteq \{A_1, ..., A_n\}$ *mit:* $\exists p \in U$ *so das gilt* $f_g(f_c(p)[Q_T]) = p$

Somit ist es anhand von Quasi-Identifiern möglich, in anonymisierten Tabellen mindestens ein Individuum eindeutig zu reidentifizieren.

Example 1. In der anonymisierten Tabelle T (2) sei das Attributset Geschlecht, Postleitzahl und Geburtstag ein Quasi-Identifier. Nimmt man nun an, dass es weltweit nur eine Person gibt die männlich, an Silvester geboren und in 76133 Karlsruhe wohnt, ist es einfach, aus der anonymisierten Tabelle ihn und insbesondere auch sein Leiden auszulesen. Eine solche Annahme ist, wie aus der amerikanischen Studie [2] zu sehen, relativ wahrscheinlich. Eine Tabelle, welche sowohl Quasi-Identifikatoren und personenspezifische Attribute enthält, ähnlich der eines Einwohnermeldeamtes, würde durch Verknüpfung anhand der Quasi-Identifikatoren auch noch die Identität dieser Person enthüllen. Somit gelangt man an diesem Beispiel zu dem Wissen, dass *Kai Traube* an Diabetes leidet.

In dieser Ausarbeitung wird der Tabelle T(2) genau ein Quais-Identifier zugeordnet, welcher die Attribute Geburtsdatum, Geschlecht und Postleitzahl umfasst. Ein Entfernen der am Quasi-Identifikator teilnehmenden Attribute insbesondere in Tabelle T (2) würde zu einem Verlust an Information führen, der die weitere Verwendung der Daten insbesondere für statistische Zwecke unbrauchbar macht. Die Problematik ist also, die individuellen Daten zu schützen und gleichzeitig den Informationsgehalt zu wahren.

Geburtstag	Geschlecht	PLZ	Krankheit
17.04.75	M	76227	Impotenz
31.07.75	M	76228	Hodenkrebs
17.01.75	M	76227	Sterilität
05.07.81	M	76133	Schizophrenie
31.12.81	M	76139	Diabetes
05.07.83	W	76133	Magersucht
31.10.83	W	76131	Magersucht

Tabelle 2. T: anonymisierte medizinische Tabelle

3 K-Anonymity

In [1] wird ein Konzept vorgestellt, welches bei Veröffentlichung von personenbezogenen Daten ein Qualitätsmaß für den Schutz vor Identifizierung eines Individuums anhand von Quasi-Identifieren darstellt. K-Anonymity gewährleistet unter Angabe von k, dass ein Individuum in den Daten von mindestens k-1 anderen ununterscheidbar bezüglich ihrer Quasi-Identifikatoren bleibt und somit anonym ist. Zitat aus Wikipedia:

> Anonymität ist das Maß der Geheimhaltung der Identität (Gleichheit, völlige Übereinstimmung, Ununterscheidbarkeit) eines Individuums bezogen auf die Gruppe, in der es agiert. Die Anonymität entspricht der Verdachtsverteilung und wächst so mit der Größe der Gruppe.

Das bedeutet für k-Anonymität: je größer k, desto größer seine Gruppe und desto anonymer also auch sicherer ist ein Individuum vor Identifizierung geschützt.

Definition 2. *Sei eine personenbezogene Tabelle* $T(A_1, ..., A_n)$
Die endliche Menge an Attributen von T $\{A_1, ..., A_n\}$ *und*
Die zur Tabelle passenden Quasi-Identifikatoren Q_T *gegeben,*
dann sagt man T *unterstützt k-Anonymity genau dann, wenn jede Wertkombination von* $T[Q_T]$ *mindestens k mal auftritt*

Example 2. Die Tabelle T'(3) ist eine mögliche Ableitung aus der Tabelle T (2) welche nun eine K=2-Anonymität unterstützt. Dabei separiert der Quasi-Identifier $Q_T = \{Geburtsdatum, Geschlecht, PLZ\}$ den Datenbestand in drei Blöcke. In T' gilt: $t1[Q_T] = t2[Q_T] = t3[Q_T], t4[Q_T] = t5[Q_T]$ und $t6[Q_T] = t7[Q_T]$ Selbst Angreifer, welche über eine komplette Datenbank aus Quasi-Identifikatoren und dazugehörigen personenspezifischen Attributen verfügen, können kein Individuum eindeutig aus dieser Tabelle bestimmen. Bei Verknüpfung mit der Tabellen T'(3) erhält man stets immer mindestens zwei mögliche Verknüpfungstuppel zu einem personenspezifischen Attribut. Dadurch wird eine eindeutige Identifizierung der Person verhindert.

	Geburtstag	**Geschlecht**	**PLZ**	**Krankheit**
t1	**.**.75	M	7622*	Impotenz
t2	**.**.75	M	7622*	Hodenkrebs
t3	**.**.75	M	7622*	Sterilität
t4	**.**.81	M	7613*	Schizophrenie
t5	**.**.81	M	7613*	Diabetes
t6	**.**.83	W	7613*	Magersucht
t7	**.**.83	W	7613*	Magersucht

Tabelle 3. T':K=2-Anonyimisierte Tabellen

	Geburtstag	Geschlecht	PLZ	Krankheit
t1	17.**.75	M	76227	Impotenz
t2	31.**.**	*	76***	Hodenkrebs
t3	17.**.75	M	76227	Sterilität
t4	05.07.8*	*	76133	Schizophrenie
t5	31.**.**	*	76***	Diabetes
t6	05.07.8*	*	76133	Magersucht
t7	31.**.**	*	76***	Magersucht

Tabelle 4. T":K=2-Anonyimisierte Tabellen

Es ist offensichtlich, dass eine unter dem Aspekt von Q_T k-anonymisierte Tabelle T auch hinsichtlich jedem Subsets von Q_T eine k-Anonymity gewährleistet. Damit liefert eine k-anonymisierte Tabelle Identifikationsschutz bei den meisten Verknüpfung mit externen Quellen, da immer mindestens k Tuppel gejoint werden müssen.

3.1 Angriffe gegen k-Anonymity

Selbst wenn alle möglichen Quasi-Identifkatoren einer zu veröffentlichenden Tabelle berücksichtigt werden, bietet k-Anonymity nicht immer optimalen Schutz. Im Folgenden werden zwei Angriffe geschildert, die bei Erstellung einer k-Anonymity beachtet werden müssen, um den angegebenen Schutz gewährleisten zu können.

Unsortiertes Verlinken Obwohl k-Anonymity einen Schutz des Individuums gegen Identifizierung durch Verknüpfung mit externen Quellen anhand von tabellenspezifischen Quasi-Identifikatoren verspricht, ist es dennoch möglich, Daten über ein Individuum auszumachen. Dies ist der Fall, wenn stattdessen die Tuppel anhand ihrer Position in der Tabelle mit einander verknüpft werden. Dieses Verfahren erfordert bestimmte Gegebenheiten in den zu verknüpfenden Tabellen. Erstmals müssen die Tabellen die gleichen Individuen enthalten und zusätzlich müssen diese in gleicher Reihenfolge vorliegen. Dies ist besonders dann der Fall, wenn von ein und derselben Tabelle mehren k-Anoymisierungen veröffentlicht wurden. Um dieses Risiko zu vermeiden, sollte eine jede zu veröffentlichende Tabelle per Zufall sortiert werden.

Example 3. Die beiden Tabellen T'(3) und T"(4), die beide aus der Tabelle T(2) hervor gehen und aus Datenschutzsicht einer k=2-Anonymity unterliegen, dienen der Veranschaulichung des Problems. Wird T'(3) mit T"(4) zeilenweise miteinander verbunden, werden somit sämtliche Daten des Tuppels t4 und t6 enthüllt. Daraufhin kann ein eindeutiger Quasi-Identifikator rekonstruiert und somit die Anonymität verhindert werden. Offensichtlich würde das einfaches Umsortieren von T"(4) ein solchen Angriff verhindern.

Komplementäre Veröffentlichung Selbst bei Sortierung vor der Veröffentlichung der Tabellen kann es zur Identifizierung der Person kommen, wenn unterschiedliche Versionen von ein und derselben Ursprungstabelle veröffentlicht werden. Wenn der Quasi-Identifier einer Person unterschiedliche Tuppel in den beiden Versionen identifiziert und die beiden Tuppelmengen sich nur in einem Tuppel gleichen, muss dieser Tuppel der Repräsentant der Person in diesen Tabellen sein. Diese Identifizierung ist möglich, da sich für nachfolgende Tabellen in Abhängigkeit der zuvor veröffentlichten Attribute der Quasi-Identifikator ändert. Deshalb müssen alle zu veröffentlichenden Tabellen von T (2) bereits existierenden K-anonymisierte Tabellen berücksichtigen.

Example 4. Kennt ein Angreifer die Quasi-Identifikatoren von *Kai Traube* und weiß, dass er sowohl in Tabelle T'(3) als auch in T"(4) vorkommen muss, dann erhält er einerseits die Tuppelmenge {t4, t5} und andererseits die Tuppelmenge {t5, t7} als mögliche Repräsentanten von *Kai's* Daten. Werden beide Mengen anhand des Attributes „Krankheit" miteinander verknüpft, erhält man nur noch die einelementige Menge mit dem Tuppel t5, welches demnach *Kai Traube* sein muss. Ein solcher Angriff kann jedoch verhindert werden, wenn bei der Veröffentlichung von T"(4) anstatt T(2) eine vorherige Veröffentlichung wie T'(3) als Grundlage benutzt wird.

3.2 Grenzen von k-Anonymity

Im Gegensatz zum vorherigen Abschnitt „Angriffe gegen k-Anonymity" wird hier das Qualitätsmaß als solches nicht angegriffen. Es wird gezeigt, dass selbst wenn der Schutz einer Person vor eindeutiger Identifizierung in einer Tabelle gewährleistet ist, es nicht für einen kompletten Schutz seiner Privatsphäre ausreicht. Die zwei im Folgenden beschriebenen Angriffe zeigen auf, dass ein Qualitiätsmaß wie k-Anonymity nicht umfassend genug ist.

Homogenity Attack K-Anonymity bietet zwar ausreichend Schutz vor eindeutiger Identifizierung eines Tuppels zu einem Individuum, allerdings reicht dies in manchen Fällen nicht aus, um alle Daten des Individuums zu schützen. Wenn die gesamte Gruppe hinter der sich ein Individuum verbirgt die gleichen Merkmale aufweist, dann verhüllt selbst eine k-Anonymity dieses Merkmal nicht.

Example 5. Annahme: Bob liest in der Bildzeitung „berühmte Schauspielerin *Kaira Knighty* im Krankenhaus" und macht sich sorgen um seine Lieblingsschauspielerin. Da aus dem Artikel nicht erkennbar ist, welches Leiden *Kaira* hat, versucht er es selbst herauszufinden. Dabei entdeckt er die 2-Anonymisierte Tabelle T'(3). Als echter Fan weiß er natürlich, wann sie geboren wurde und wo sie lebt. Somit findet er heraus, dass ihre Daten entweder in Tuppel t6 oder t7 hinterlegt wurden. Da beide Tuppel allerdings die selbe Krankheit aufweisen, weiß er, dass sie an Magersucht leiden muss.

K-Anonymity und spezifisches Hintergrundwissen Wie aus dem Homogenity Attack hervorgegangen ist, schützt k-Anonymity nicht zwingend die sensitiven Attributwerte eines Individuums.

Definition 3. *Man spricht von einem sensitiven Attribut, wenn dessen Werte bei Zuordnung zu einem Individuum diesem schadet oder dessen Privatsphäre beeinträchtigt.*

Wenn zu einem spezifischen Q_T mehrere Tuppel mit unterschiedlichen sensitiven Attributwerten zugeordnet werden, ist es trotzdem durch Ausschluss von sensitiven Attributwerten möglich, auf die richtigen zu schließen. Ein solcher Angriff ist denkbar, wenn der Angreifende über Hintergrundwissen in Bezug auf das Opfer verfügt.

Example 6. Jedes Mal, wenn *Till Schwäger* sich für eine ernsthafte Rolle interessiert, bekommt er zwar ein Stück Kuchen, aber keinen Auftrag, warum? Jeder gute Produzent weiß, dass *Till* 83 geboren wurde und dass er in der Innenstadt von Karlsruhe wohnt. Des weiteren ist bekannt, dass er einen längeren Krankenhausaufenthalt hatte. Anhand dieser Daten ist es nun möglich, Till Schwäger in der dem Krankenhaus zuzuordnenden Tabelle T' (3) als Tuppel t4 oder t5 zu identifizieren. Es ist somit nicht bekannt, ob er Schizophren ist oder unter Diabetes leidet. Deshalb bekommt Herr *Schwäger* jedesmal einen Kuchen und die in der Situation recht harmlos wirkende Frage „möchten sie dieses Stück Kuchen oder sind sie Diabethiker" gestellt. Da er gerne Kuchen isst und kein Diabetes hat, nimmt er jedes Mal das Stück Kuchen an. Damit kann der Produzent allerdings ausschließen, dass er Diabetiker ist und schlussfolgert somit, dass *Till* an Schizophrenie leidet. Da keiner ernsthafte Rollen an schizophrene Schauspieler verteilt, wird Till stets nur mit einem Stück Kuchen abgespeist.

4 L-Diversity

Es wurde gezeigt, dass k-Anonymity keinen ausreichenden Schutz der sensitiven Attribute gegen Identifizierung mit einer realen Person bietet. Bei K-anonymen Tabellen ergeben sich zwei Möglichkeiten die Information über das sensitive Attribut einer Person herauszufinden. Erstens über „positive disclosure", bei der mit einer hohen Sicherheit der sensitive Attributwert vorhergesagt wird und zweitens durch „negative disclosure", welches die Zahl der in Frage kommenden sensitiven Attributwerte, durch Ausschluss der mit hoher Wahrscheinlichkeit nicht in Frage kommender Attribute, reduziert. Im Falle eines positiven Ausschlusses ist der Schutz der Privatsphäre nicht mehr gewährleistet, was es nun zu verhindern gilt.

4.1 Prinzip der l-Diversity

Das Paper[7] von *Ashwin M. Johannes G.und Daniel K.* behandelten dieses Problem und definierten ein Sicherheitsmaß für den Schutz der sensitiven Attribute.

Dafür wird eine Tabelle in sensitive und nicht sensitive Attribute unterteilt. Um das Prinzip der l-Diversity besser erläutern zu können, wird im Folgenden davon ausgegangen, dass alle nicht sensitiven Attribute den Quasi-Identifier bilden und die Tabellen lediglich ein sensitives Attribut besitzen. Da vorab nicht jedes Hintergrundwissen bekannt sein kann, gibt es keinen in der Praxis umsetzbaren optimalen Schutz gegen Ausschluss sensitiver Attribute. L-Diversity bietet daher nur ein Schutzmaß vor „positiven disclosure“. Hierbei wird die Verbindung eines sensitiven Attributes mit einer Person geschützt, in dem dieses mindestens in einer Menge von l anderen sensitiven Attributen versteckt wird. Ein Angreifer benötigt somit mindestens l-1 Hintergrundwissen, um durch genügend Ausschluss der falschen sensitiven Attribute auf das richtige schließen zu können.

Definition 4. *Gegeben sei T* eine K-anonymisierte Tabelle von T und q*-Block eine Menge an Tuppeln, die sich anhand von Q_{T*} nicht unterscheiden Ein q*-Block ist l-diverse, wenn er mindestens l „gut repräsentierte“ Werte für die sensitiven Attribute besitzt. Eine Tabelle besitzt l-Diversity, wenn jeder q*-Block l-diverse ist.*

„Gut repräsentiert“ hängt in diesem Kontext von der konkreten Instanzierung von l-Diversity ab.

4.2 konkrete l-Diversity Instanzen

L-Diversity bietet fünf mögliche Instanzierungen um „gut repräsentiert“ zu definieren, welche im Folgenden erklärt werden. Allerdings wird im Rahmen dieser Ausarbeitung nur auf Entropy l-Diversity detailliert eingegangen.

Entropy l-Diversity

Definition 5. *Gegeben sei T* eine K-anonymisierte Tabelle von T,*
S die Menge an sensitiven Attributwerten und
q-Block eine Menge an Tuppeln, die sich anhand von Q_{T*} nicht unterscheiden*
Eine Tabelle T besitzt Entropy l-Diversity wenn für jeden q*-Block gilt:*
$-\sum_{s \in S} P_{(q*,s)} * log(P_{(q*,s)}) \geq log(l)$,

wobei $P_{(q*,s)} = \frac{n_{(q*,s)}}{\sum_{s' \in S} n_{(q*,s')}}$, hierbei sei $n_{(q*,s)}$ die Anzahl sensitiver Attributwerte s in einem q-Block.*

Aus dieser Definition folgt, dass jeder q*-Block mindestens l unterschiedliche sensitive Werte aufweist. Tabelle T’ (3) entspricht wegen dem q*-Block mit Tuppel t6 und t7 einer Entropy-1-Diversity und schützt somit das sensitive Attribut Krankheit vor l-1 Hintergrundwissen; also gar keinem. Je höher l gewählt wird, desto mehr Schutz wird den sensitiven Attributen geboten. Unter der in dieser Ausarbeitung getroffenen Vereinfachung, dass alle nicht sensitiven Attribute den Quasi-Identifikator bilden, kann l maximal den Wert von k annehmen.

Example 7. T* (5) zeigt eine Tabelle, die zwei q*-Blöcke besitzt und einer Entropy-2.8-Diversity genügt. Mit l größer als eins ist trivialer Weise darauf ein homogenity Attack nicht mehr möglich. Selbst ein Angriff auf die Privatsphäre von Till Schwäger muss dieses Mal neben dem Ausschluss der Diabeteserkrankung zusätzlich Magersucht elemenieren, bevor eindeutig Schizophrenie als sensitives Attribut identifiziert werden kann. In diesem Fall muss ein Angreifer also immer mindestens zwei sensitive Attribute mit hoher Wahrscheinlichkeit ausschließen, bevor er zu einem „positiven disclosure“ kommt.

Geburtstag	Geschlecht	PLZ	Krankheit
..75	M	76227	Impotenz
..75	M	76228	Hodenkrebs
..75	M	76227	Sterilität
..8*	*	7613*	Schizophrenie
..8*	*	7613*	Diabetes
..8*	*	7613*	Magersucht
..8*	*	7613*	Magersucht

Tabelle 5. T*: K=3-anonymisierte Entropy-l=2.8-divserse Tabelle

Recursive(c,l)-Diversity Hierbei wird darauf geachtet, dass keines der sensitiven Attribute in jedem q*-Block zu häufig vorkommt. Je kleiner c desto gleich verteilter sind die sensitiven Attribute. L-1 entspricht der Anzahl an sensitiven Attributwerten, die in einem q*-Block ausgeschlossen werden können, bevor ein Attributwert zu häufig vorkommt.

Positive Disclosure-Recursive(c,l)-Diversity Es ist möglich, dass einige positive Enthüllungen des sensitiven Attributes erlaubt sind, da sie fast jedes Individuum besitzt oder keine Beeinträchtigung der Privatsphäre bewirkt, wenn sie bekannt würde. Ein Beispiel dafür wäre das sensitive Attribut „Krankheit“, das den Attributwert „gesund“ enthalten würde. Dies wird in der positive disclosure-rekursive(c,l)-Diversity (PD-R-(c,l)-Diversity) berücksichtigt, indem ein zu häufiges Vorkommen eines unkritischen sensitiven Attributwertes erlaubt wird.

Negative/Positive Disclosure-Recursive(c1,c2,l)-Diversity Ausgehend von PD-R-(c,l)-Diversity wird hier anhand von c2 ein prozentueller Gehalt an sensitiven Attributwerten in jedem q*-Block definiert, welche nicht ohne Weiteres einen Ausschluss erlauben.

Multiple Sensitive Attributes Wie der Name schon sagt beschäftigt sich diese Instanzierung von l-Diversity damit, einen ausreichenden Schutz vor Veröffentlichung mit mehreren sensitiven Attributen zu gewährleisten. Dabei wird darauf geachtet, dass die sensitiven Attribute einer Person auch dann geschützt bleiben, wenn bereits ein anderes sensitives Attribut identifiziert worden ist.

4.3 Grenzen der l-Diversity

Tabelle T* (5) zeigt, dass das Maß Entropy-l-Diversity die sensitiven Daten einer Person in einer Gruppe mit mindestens l unterschiedlichen sensitiven Attributen vor Angreifern schützt. Da lediglich die eindeutige Zuweisung eines sensitiven Attributes zu einer Person im Schutzmaß wiedergespiegelt wird, bietet auch dieses Konzept keine Gewährleistung vor allgemeinem Informationsgewinn eines Angreifers bezüglich der sensitiven Attribute. Zwei Angriffe werden dies im Folgenden erläutern.

Skewness Attack Wenn eine zu veröffentlichende Tabelle einen sensitiven Attributwert besitzt, den 99% der Bevölkerung besitzen und nur 1% nicht, dann würde ein q*-Block, welcher nur zu 50% diesen Wert besitz hingegen zu 50% einen anderen, zwar Entropy-L=2-diversity und diversen recursive(c,2)-diversity schutzmaßen genügen, einem Schutz vor Infromationsgewinn jedoch nicht. Jedes Individuum, dass sich in dieser Gruppe befindet, hat demnach eine deutlich höhere Chance anormal verglichen mit der Allgemeinheit zu sein.

Example 8. Angenommen, dass sensitive Attribut könnte nur 2 Werte annehmen. Eine Person ist krank oder ein Person ist nicht krank und nur ein Prozent der gesamten Bevölkerung ist krank und 99% sind gesund. Wenn Alice nun ein Vorstellungsgespräch hat und Bob, ihr zukünftiger Chef, stellt fest, sie befindet sich in einer Gruppe, die zu 50% krank ist, dann wirkt sich das möglicherweise auf ihre Einstellung aus, obwohl aus l-Diversity Sicht ein guter Schutz gegeben war.

Similarity Attack Dieser Angriff ähnelt dem Homogenity Attack und ist möglich, wenn die Gruppe in der ein Individuum anonymisiert wurde, eine Gemeinsamkeit hat. Dabei kann zwar jedes Individuum aus dieser Gruppe einen anderen Wert des sensitiven Attributes besitzen und somit einem l-Diversity-Maß genügen, allerdings können sich auch unterschiedliche Attributwerte ähneln. Ist das in der gesamten Gruppe der Fall, so kann das sensitive Attribut eines Individuums durch einen Angreifer zumindest kategorisiert werden. L-Diversity bietet somit keinen ausreichenden Schutz.

Example 9. Karl Feldlager und seine langjährige Freundin Alice versuchen seit zwei Jahren Kinder zu zeugen. Alice bekommt nun zufällig die 2.8-Diverse-Tabelle (5) des Krankenhauses, in dem auch *Karl* war, zu Gesicht. Sie erkennt anhand des Geburtsjahres und der Postleitzahl, dass ihr Freund wohl einer der oberen drei Tuppel sein müsste und folglich gar nicht in der Lage ist, Kinder zu

zeugen. Obwohl diese Tuppel drei unterschiedliche sensitive Attributwerte enthalten, kann diese Gruppe an Tuppeln in die Kategorie der Zeugungsunfähigen eingestuft werden. Einerseits ein Trennungsgrund andererseits ein Beweis, dass der Schutz vor Kategorisierung des sensitiven Attributes nicht zu vernachlässigen ist.

5 T-Closeness

Die Arbeit [8] von *Ninghui Li Tancheng Li* und *Suresh V.* beschäftigt sich mit einem neuen gegenüber l-Diversity verbesserten Maß an Sicherheit. Es berücksichtigt den nicht zu verhindernden Gewinn an Wissen eines Angreifers bei Betrachtung aller sensitiven Attributwerte in der gesamten Distribution. T-Closeness stellt ein Maß für minimalen Wissensgewinn, der durch Betrachtung eines q*-Blocks im Vergleich zur gesamten Distribution entsteht, dar. Das bedeutet auch, dass jede, anhand des Quasi-Identifikators ununterscheidbare, Gruppe an Individuen, hinter der sich eine Person anonymisiert, durch das t-Closeness definierte Maß kaum von jeder anderen Gruppe bezüglich ihrer sensitiven Attributwerte unterschieden werden kann. Somit sind die Daten einer Person besser in seiner anonymisierenden Gruppe geschützt, als dies bei bei l-Diversity der Fall war, da diese Gruppe kaum mehr Information preisgibt als die gesamte Distribution.

Definition 6. *Ein q*-Block besitzt t-Closenenss, wenn die Distanz zwischen den zu veröffentlichenden sensitiven Attributen dieses Blocks zur gesamten Tabelle nicht mehr als ein Grenzwert von t beträgt. Eine Tabelle besitzt t-Closeness, wenn alle q*-Blöcke von ihr t-Closeness besitzen.*

Ein nicht zu vernachlässigendes Problem stellt allerdings die Distanzmessung zwischen den Attributen dar. Einfache Verfahren, wie die variational Distanz, beziehen die semantische Ähnlichkeit zweier Werte zueinander nicht mit ein.
Variational Distanz: $D[P,Q] = \sum_{i=1}^{m}(1/2|p_{(i)} - q_{(i)})$
Es ist also ein spezielles Distanzmessungsverfahren von Nöten. Die Earth Mover's distanz (EMD)[11] verspricht eine Lösung für das Problem zu sein. Es basiert auf der minimalen Arbeit, die zu verrichten ist, um eine Distribution in eine andere umzuwandeln.

Definition 7. *Gegeben sei* $P = (p_1, ..., p_m)$,
$Q = (q_1, ..., q_m)$,
d_{ij} *die Grund Distanz zwischen Element* p_i *und* q_j.
f_{ij} *ist die minimale Masse, die transportiert werden muss um* p_i *in* q_i *zu verwandeln.*
EMD sei dann die gesamte Arbeit die verrichtet werden muss
$D[P,Q] = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij}f_{ij}$
Unter den Bedingungen:

1. $f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq m$

2. $p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ij} = q_i \qquad 1 \leq i \leq m$

3. $\sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1$

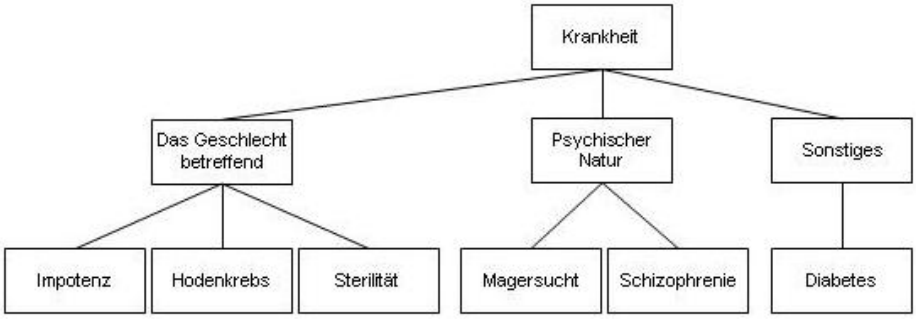
Wenn die Grunddistanz normalisiert zwischen Null und Eins vorliegt folgt aus der Bedingung 1 und 3, dass die EMD Distanz auch zwischen Null und Eins liegt und somit ein einheitliches Maß für t von t-Closeness bestimmt werden kann. Bisher wurden jedoch nur numerische Attributwerte berücksichtigt. Bei kategorischen Attributwerten muss zunächst noch eine Umrechnungsformel definiert werden, mit der sich Distanzen berechnen lassen. Der einfache Ansatz, jedem unterschiedlich kategorischen Wertepaar die Distanz Eins und jedem gleichem Wertepaar die Distanz Null zuzuordnen, lässt wiederum semantische Ähnlichkeiten außer Acht. Ein gängiges Verfahren ist deshalb, die Distanz zweier kategorischer Attributwerte anhand von Generalisierungsschritten zu bewerten. Dabei sind immer soviel Generalisierungen nötig, um den einen Attributwert in die gleiche Domäne wie den anderen zu transformieren. In Figur (1) kann demnach Hodenkrebs zu Impotenz in einem Generalisierungsschritt zu Geschlechtskrankheit transformiert werden, wohingegen Diabetes und Impotenz erst über zwei Generalisierungen die gemeinsame Domäne Krankheit besitzen. Normalisiert man diese Abstandsmaße auf einen Wertebereich zwischen Null und Eins kann nun die t-Closeness von z.B. Tabelle T* (5) ermittelt werden. Im Gegensatz zu k-Anonymity und l-Diversity bietet hier eine kleinere Variable t einen höheren Privatsphärenschutz. Die rechte Tabelle (6) zeigt eine bezüglich dem Sicherheitsmaß t-Closeness bessere Tabelle verglichen zu T*. Allerdings wurde dafür das Konzept Anatomy verwendet, da die Tabelle T für optimale Werte von t-Closeness nur durch Generalisierung sämtlicher Attribute des Quasi-Identifikators erreicht werden könnte. Aus der rechten Tabelle ist ersichtlich, dass der Similarity Attack die Attributwerte von Krankheit nicht weiter kategorisieren kann. Eine Kategorisierung des sensitiven Attributes ist bei guten t-Closeness geschützten Tabellen nur dann der Fall, wenn eine Ähnlichkeit aller sensitiven Attribute zueinander vorliegt. Vor einem solchen Informationsgewinn könnte in dieser Konstellation allerdings kein Konzept Schutz bieten.

6 Probleme der vorgestellten Konzepte

Alle diese Konzepte haben eines gemeinsam: sie müssen zum Erstellen von anonymisierenden Gruppen Informationen aus den Quasi-Identifikatoren löschen. Dabei kommen vier mögliche Verfahren zum Einsatz:

1. Dataswaping: Attributwerte werden zwischen den Tuppeln ausgetauscht
2. Ading Noise: Fehlinformationen werden in die Attributwerte eingerechnet
3. Generalisierung: die Attributwerte werden kategorisiert
4. Supression: Attributwerte werden eliminiert und ignoriert

Abbildung 1. eine mögliche Wertebereichshierarchie von Krankheit

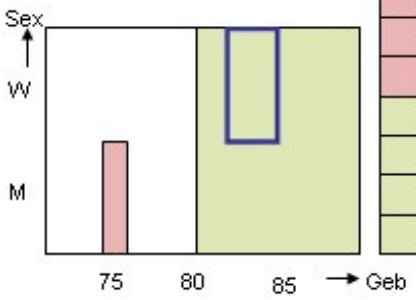


In dieser Ausarbeitung wurde jedoch nur Generalisierung wie in Tabelle T' (3) oder Supression wie in Tabelle (5) auf dem Attribut „Geschlecht“ durchgeführt. Eine optimale k-Anonymity versucht die Orginaldaten so wenig wie möglich zu manipulieren. Tabelle T" (4) ist ein eindrucksvolles Beispiel für eine schlechte Anonymisierung. Ein erwähnenswerter Fakt ist, dass die Berechnung einer optimalen k-Anonymity ein NP-hartes Problem darstellt und die Berechnung der Gruppen unter Berücksichtigung optimaler l-Diversity oder t-Closeness das Problem nicht mildern werden. Zudem verlieren die Tabellen durch diese Verfahren an Information, was die analytischen Anfragen auf solche Tabellen stark verfälscht. Wie stark die Informationen manipuliert wurden, veranschaulicht folgendes Beispiel:

Example 10. Die Tabelle T* (5) zeigt eine stark generalisierte Tabelle, in der nur noch zwei anhand der Quasi-Identifier unterscheidbare Blöcke anzutreffen sind. Die Grafik (2) zeigt links eine zweidimensionale Darstellung der beiden Blöcke, wobei hier zur besseren Veranschaulichung vom Attributwert „Postleitzahl“ und „Krankheit“ abstrahiert wurde und auf der rechten Seite noch einen Ausschnitt aus T*. Man sieht, dass der rechte grün hinterlegte Block alle Tuppel zeigt, welche zwischen den Jahren 1979 und 1990 geboren wurden und entweder männlich oder weiblich sind. Der linke rötliche Block hingegen beschreibt männliche, im Jahr 1975 geborene Tuppel. Eine statistische Anfrage wie Beispielsweise: “Wie viele weibliche Magersüchtige gab es im Jahre 1983?“ entspricht demnach in der zweidimensionalen Tabelle dem blau umrandeten Viereck. Aus Tabelle T (1) ist ersichtlich, dass es ursprünglich zwei weibliche Magersüchtige im Jahre 1983 gab. Nach der Generalisierung allerdings gibt es für die Anfrage nur noch zwei Datenpunkte, die bezüglich ihrer Quasi-Identifikatoren nicht weiter unterschieden werden können. Die Anfrage kann somit nur herausfinden, dass in dem grünen Block zwei Magersüchtige sind, nicht aber wie viele Magersüchtige sich unterhalb des blauen Rechtecks befinden. Eine exakte Antwort kann somit nicht bestimmt, sondern lediglich nähreungsweise ermittelt werden. Da das blaue Rechteck nur 1/20 des grünen Rechtecks abdeckt und sich im grünem Bereich zwei Magersüchtige befinden, würde das einer Wahrscheinlichkeit von 1/10 Ma-

gersüchtigen in der Anfrage entsprechen. Die Anfrage erhält somit auf T* ein stark verfälschtes Ergebnis gegenüber der selben Anfrage auf der Orginaltabelle.

Abbildung 2. links: zweidimensionale Darstellung von T* und rechts: Ausschnitt von T*



Geburts-Jahr	Geschlecht	Krankheit
75	M	Impotenz
75	M	Hodenkrebs
75	M	Sterilität
8*	*	Schizophrenie
8*	*	Diabetes
8*	*	Magersucht
8*	*	Magersucht

6.1 Anatomy

In [5] wurde ein Verfahren beschrieben, welches unter Verwendung jeder zuvor erläuterten Qualitätsmaße nur geringfügig weniger Sicherheit bietet bei gleichzeitig höherem Wiederverwendungswert der Daten. Bei diesem Verfahren wird auf eine Veränderung der Quasi-Identifikatorwerte zum Erhalt von ununterscheidbaren Gruppen verzichtet. Stattdessen werden die Quasi-Identifikatoren und die sensitiven Daten von einander getrennt und in zwei unterschiedlichen Tabellen veröffentlicht. Dabei wird jedem Tuppel der Orginaltabelle ein „Join-Attribut“ angehängt, welches eine Kennung seiner Gruppe widerspiegelt. Bei Aufsplittung in die zwei zu veröffentlichenden Tabellen erhält jede Tabelle diese Gruppenzugehörigkeits-ID. Eine k-Anonymity kann also gewährleistet werden, wenn jedem Tuppel eine Gruppenidentifikation zugeteilt wird und jeder Gruppe mindestens k Tuppel angehören. Bei Verknüpfung der Quasi-Identifikator-Tabelle mit der sensitiven Tabelle erhält man mindestens k sensitive Attribute und hat somit sein sensitives Attribut in der Menge von k-1 falschen Attributen anonymisiert. Es ist ersichtlich, dass mit dieser Methode auch die Konzepte l-Diversity und t-Closeness umgesetzt werden können, wie die Beispiel Tabellen (6) für t-Closeness zeigen. Allerdings enthalten die Tabellen nun die möglicherweise einzigartigen Kombinationswerte des Quasi-Identifikators. Damit gibt es die Information über das Vorhandensein eines Individuums in einer Tabelle preis (siehe [2]), die als Grundvoraussetzung der meisten aufgezeigten Angriffen diente.

Geburtstag	Geschlecht	PLZ	GruppenID
17.04.75	M	76227	2
..75	M	76228	1
..75	M	76227	2
..8*	*	7613*	2
..8*	*	7613*	1
..8*	*	7613*	2
..8*	*	7613*	1

Krankheit	GruppenID
Impotenz	2
Hodenkrebs	1
Sterilität	2
Schizophrenie	2
Diabetes	1
Magersucht	2
Magersucht	1

Tabelle 6. T*: K=3-anonymisierte Entropy-l=2.8-divserse 0,29-Closeness

7 Fazit

Anhand von Beispielen mit personenbezogenen Daten wurde die Notwendigkeit des Schutzes der Privatsphäre erläutert und nacheinander verschiedene Konzepte aufgezeigt, die ein Maß an Sicherheit in zu veröffentlichenden Tabellen gewährleisten sollen. K-Anonymity besticht hierbei durch seine Einfachheit und einem ausreichenden Schutz des Individuums vor eindeutiger Identifizierung innerhalb einer Veröffentlichung. Sie weist jedoch Schwächen gegen Angreifer mit Hintergrundwissen und gegenüber Angriffen auf die sensitiven Attribute auf. L-Diversity geschützte Tabellen besitzen selbst wieder eine k-Anonymity und schützen zusätzlich die sensitiven Attribute. Jede Gruppe, in der sich ein Individuum anonymisiert, besitzt dabei mindestens l unterschiedliche sensitive Attribute. Unberücksichtigt blieb hier, dass auch unterschiedliche sensitive Attributwerte semantische Ähnlichlkeiten aufweisen. Diesen Aspekt beachtet das t-Closeness-Konzept, welches zwar das bisher höchste Maß an Sicherheit gewährleistet, in der Umsetzung aber am meisten Aufwand benötigt. Unter Anderem muss hier ein Distanzmaß für kategorische Attribute gefunden werden, wobei zu bedenken ist, dass eine hierarchische Attribuwertgeneralisierung nicht immer eindeutig ist. Ob dieses Sicherheitsmaß jedoch vertraulich genug ist, um alle Informationen über die Privatsphäre zu schützen, wird sich erst zeigen müssen. Da dieses Konzept noch neu ist, ist derzeit noch kein Angriffsszenario bekannt. Das Konzept von Anatomy schließlich stellt aus Sicht der statistischen Datenerfassung eine Verbesserung zu den vorgestellten Konzepten dar und kann jedes dieser Qualitätsmaße zum Schutz der Privatsphäre einsetzen bei gleichzeitig besserer Ausnutzung der Daten für statistische Zwecke. Einziger Nachteil von Anatomy ist, dass ein Individuum nun eindeutiger (z.B. 87% der amerikanischen Bevölkerung) einer Tabelle zugeordnet werden kann. Meiner Meinung nach ist derzeit die Verwendung von Anatomy mit dem Sicherheitsmaß t-Closeness zu empfehlen.

Literatur

[1] L. Sweeney, K-anonymity: A model for protecting privacy, international journal of uncertainity, fuzziness and knowledge-based systems 10(5):557 - 570, 2002

[2] L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population Population, Carnegie Mellon University, laboratory for international data privacy LIDAPWP4, 2000

[3] R.J. Bayardo and R. Agrawal, Data Privacy through optimal k-Anonymization, proceedings of the 21st iInternational conference on data engineering:217 - 228, 2005

[4] A. Meyerson and R. Williams, On the Complexity of optimal k-Anonymity, proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems:223 - 228, 2004

[5] X. Xiao and Y. Tao, Anatomy: simple and effective privacy preservation, proceedings of the 32nd international conference on very large data bases:139 - 150, 2006

[6] C.C. Aggarwal, On k-anonymity and the curse of dimensionality, proceedings of the 31st international conference on very large data bases:901 909, 2005.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, L-diversity: Privacy beyond k-anonymity, proceedings of the 22nd international conference on data engineering:24 - 36, 2006

[8] N. Li and T. Li, t-closeness: Privacy beyond k-anonymity and l-diversity, proceedings of the 23rd international conference on Data Engineering:106 - 115, 2007.

[9] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, international journal of uncertainity, fuzziness and knowledge-based systems 10(6):571 588, 2002

[10] T. Dalenius, Finding a needle in a haystack or identifying anonymous census record, journal of official statistics 2(3):329 - 336, 1986.

[11] Y. Rubner, C. Tomasi and L.J. Guibas, The earth movers distance as a metric for image retrieval, proceedings of the 1998 international conference on computer vision:59 - 66, 1998