Ideen, Ansätze, Fragen

- Formalisierung von k-Anonymität, wenn man annimmt einen Wert der Ausgangstabelle zu kennen (Anmerkung: Schlechte Idee, da man somit nur in die Äquivalenzklasse navigiert und keine Eindeutigkeit entsteht. Deswegen gibt es ja diese Äquivalenzklassen)
- Überlegen, was ist die optimalste Vairante der Anonymisierung
- (Alte) Bekannte Anonymisierungsmethoden: Datfly, μ-Argus, k-Similar (Bilden eine gewisse Grundlage für den k-Anonymisierungsansatz)
- Idee zur Evaluierung: Eine Spalte nicht mitanonymisieren und gucken wie viel man dann anhand des Wissens der Spalte lösen kann
- Frage: Wie sieht anonymisierung aus, wenn es mehrere Tabellen mit ähnlichen Einträgen gibt?
- Frage: Ist es möglich die Anonymität einzelner Datensätze zu bewerten?

Inhaltsverzeichnis

1 Literaturrecherche

1.1 Grundlegendes

1.1.1 Wichtige Begrifflichkeiten

Diese Arbeit befasst sich hauptsächlich mit Anonymisierung und wie man diese evaluieren kann. Dabei denkt man auch häufig an den Begriff der Pseudonymisierung, welcher im heutigen Cyberspace oft fälschlicherweise als Anonymisierung interpretiert wird. Zum Verständnis der Begrifflichkeiten und Unterschiede wird dies im Folgenden definiert:

Definition 1.1

- **Anonymisierung**: Dabei werden sensible personenbezogene Daten derartig verändert, dass diese nicht mehr oder nur mit hohem Kostenaufwand dem ursprünglichen Individuum eindeutig zugeordnet werden können
- **Deanonymisierung**: Damit wird die gezielte Aufhebung vorher durchgeführter Anonymisierung von Daten beschrieben
- **Pseudonymisierung**: Dabei wird das Identifikationsmerkal eines Individuums durch ein Pseudonym ersetzt, sodass die Feststellung der Identität nicht mehr oder nur mit hohem Kostenaufwand möglich ist

Seien die zu anonymisierenden Daten in einer Tabelle eingetragen. Jede Zeile stellt einen Datensatz dar, welcher aus Attributen besteht. Somit stellt jede Spalte ein Datensatzübergreifendes Attribut dar. Diese lassen sich in drei Kategorien einteilen:

Definition 1.2 (Klassifikation von Attributen)

- **Identifikatoren**: Damit ist die direkte Identifikation von Individuen möglich. Beispiele: Auto-Kennzeichen, Ausweisnummer, Matrikelnummer, etc.
- **Quasi-Identifikatoren**: Diese Attribute ermöglichen keine direkte Identifikation von Individuen, aber eine Kombination aus mehreren Quasi-Identifikatoren schon. Beispiel: Geburtstag, Postleitzahl, Geschlecht, Beruf, etc.
- **Sensible Attribute**: Darin sind schützenswerte Informationen gespeichert. Mittels Anonymisierung soll die Zuordnung dieser Daten zu einem Individuum verhindert werden. Beispiele: Krankheiten, Steuerinformationen, etc.

1.1.2 Erzeugung anonymisierter Datensätze

Die Methoden zur Anonymisierung sind meist invdividualisiert und auf den Kontext der Daten zugeschnitten. Zudem werden bis heute oftmals keine standardisierten Verfahren verwendet. Dennoch können die verwendeten Ansätze in fünf Kategorien eingeteilt werden:

Definition 1.3 (Klassifikation von Anonymisierungsmethoden)

- **Generalisierung**: Ersetzung von Attributswerten durch allgemeinere Werte
- **Informationsunterdrückung**: Löschung von Attribute aus dem Datensatz
- **Anatomisierung**: Verschleierung der Beziehungen zwischen den Attributen
- **Permutation**: Verschleierung der Beziehung zwischen Quasi-Identifikatoren und Sensiblen Attributen durch Vertauschung der Attributswerte
- **Perturbation**: Ersetzung von Attributswerten durch Rauschen oder synthetische Werte

1.2 k-Anonymität

Mit dem k-Anonymitätsansatz kann die Anonymisierung von Datensätzen evaluiert werden. Die Grundidee ist zu Zeigen, dass beim Reanonymisieren jedes einzelne Individuum von mindestens

$(k-1)$ anderen Individuen ununterscheidbar ist.

1.2.1 Funktionsweise

Die k-Anonymität wird anhand der anonymisierten Tabelle bestimmt. In der Literatur wird dies folgendermaßen definiert:

Definition 1.4 (k-Anonymität [4])

> Sei $RT(A_1, ..., A_n)$ eine Tabelle mit den Attributen A_i und QI_{RT} die darin enthaltenen Quasi-identifikatoren. RT erfüllt die k-Anonymitätseigenschaft genau dann wenn die Werte jeder Quasi-identifikatorenfolge aus $RT[QI_{RT}]$ mindestens k-mal in $RT[QI_{RT}]$ vorkommen.

Vereinfacht kann man dies mit dem Konzept von Äquivalenzklassen darstellen. Eine zu anonymisierend Tabelle enthält Quasi-identifikatoren als Spalten und Zeilen welche sowohl den Quasi-Identifiaktoren aber auch anderen Attributen werden zuweisen. Beim Anonymisieren werden die Einträge der Quasi-Identifiaktoren verändert und es entsteht eine neue Tabelle. Fasst man in dieser Tabelle alle Zeilen zusammen, welche die exakt gleichen Werte für alle Quasi-Identifikatoren haben, so erhält man Äquivalenzklassen. Die Mächtigkeit der kleinsten erhaltenden Äquivalenzklasse definiert schließlich das k der k-Anonymisierung. Anhand folgendem Beispiel kann dies noch einmal genauer nachvollzogen werden:

Beispiel 1.5. (k-Anonymisierung mittels Generalisierung)

Gegeben sei eine Tabelle als nicht anonymisierte Datenbank für medizinische Daten:

Name	Alter	Geschlecht	PLZ	Krankheit
Alice	21	W	85051	Grippe
Bob	37	M	91042	Krebs
Carol	55	M	85073	Krebs
Dan	58	M	85053	Grippe
Eve	27	W	85014	Pest
Frnk	36	M	91055	Krebs
Grace	22	W	85051	Pest

Verwendet man nun das Prinzip der Generalisierung zur Anonymisierung ergibt sich folgende Tabelle

Name	Alter	Geschlecht	PLZ	Krankheit
*	[20,30]	W	85***	Grippe
*	[30,40]	M	91***	Krebs
*	[50,60]	M	85***	Krebs
*	[50,60]	M	85***	Grippe
*	[20,30]	W	85***	Pest
*	[30,40]	M	91***	Krebs
*	[20,30]	W	85***	Pest

Daraus kann man drei Äquivalenzklassen (rot, blau, grün) bilden:

Name	Alter	Geschlecht	PLZ	Krankheit
*	[20,30]	W	85***	Grippe
*	[20,30]	W	85***	Pest
*	[20,30]	W	85***	Pest
*	[30,40]	M	91***	Krebs
*	[30,40]	M	91***	Krebs
*	[50,60]	M	85***	Krebs
*	[50,60]	M	85***	Grippe

Da jede Äquivalenzklasse zwei oder mehr Elemente enthält, gewähleistet die durch Generalisierung anonymisierte Tabelle 2-Anonymität. Zudem erkennt man an dem Beispiel auch eine Schwachstelle von k-Anonymität. Die Klasse mit Bob und Frank hat für alle Elemente den selben Wert für das sensible Attribut. Dies ermöglicht somit Homogenitätsattacken.

In den folgenden Unterkapiteln wird nun genauer auf die Schwächen der k-Anonymisierung eingegangen. Anschließend werden in den darauffolgenden Kapiteln mit l-diversity und t-closeness zwei Erweiterungen vorgestellt, deren Ziel die Überwindung ausgewählter Schwachstellen ist.

1.2.2 Schwächung der Anonymisierung durch fehlende Umsortierung

Bleibt die Reihenfolge der Zeilen der zu ursprünglichen Tabelle beim Anonymisieren erhalten, so schwächt dies die Anonymisierung im Allgemeinen. Daher ist es für die Evaluation mittels k-Anonymität wichtig, dass die Zeilen während des Prozesses randomisiert vertauscht werden [4].

1.2.3 Problem bei unterschiedlich anonymisierten Tabellen

Im obrigen Beispiel wurden alle Quasi-Attribute zur Anonymisierung verändert. Ist dies nicht der Fall, so können abhängig von einer Tabelle mehrere unterschiedlich anonymisierte Tabellen erstellt werden. Durch geschickter Kombination dieser Tabellen können Teile Rerandomisiert werden. Eine Lösung zur Erstellung mehrerer anonymisierter Tabellen abhängig von einer zugrundeliegenden Tabelle ist es, bereits veränderte Quasi-Attribute bei der Erstellung zusätzlicher Tabellen weiterhin entsprechend zu verändern und gegebenfalls nur zusätzliche Attribute als Quasi-Attribute hinzuzufügen und zu verändern. Die andere Möglichkeit ist es, prinzipiell alle Attribute (bis auf sensible Attribute und Identifikatoren) als Quasi-Attribute beim Anonymisieren zu verändern, sodass dieses Problem garnicht erst auftritt [4].

1.2.4 k-Anonymität und dynamische Tabellenanonymisierung

Datensätze können dynamisch vergrößert, verkleindert und verändert werden. Dadurch entsteht das gleiche Problem wie bei unterschiedlich anonymisierten Tabellen. Die Lösungen für das Problem sind dementsprechend ähnlich [4].

1.3 l-Diversity als Erweiterung von k-Anonymität

Ein Problem, welches bei k-Anonymität auftreten kann und nicht mit Hilfe dieses Ansatzes gelöst werden kann, sind homogene Äquivalenzklassen. Dabei haben alle Datensätze den gleichen Eintrag für das sensible Attribut. Im obrigen Beispiel entspricht dies der blauen Äquivalenklasse.
Ein weiteres Problem ist das Vorhandensein von Zusatzwissen, was auch „Background Knowledge Attack“ genannt wird. Enthält eine Äquivalenzklasse nur augewählte sensible Attribute, so kann man durch Zusatzwissen gewisse Datensätze ausschließen und damit die Anonymität umgehen. Im obrigen Beispiel entspricht dies einem Angreifer, welcher weiß, dass seine Freundin in der roten Äquivalenzklasse ist. Nehme man nun an, dass der Angreifer weiß, dass die Freundin keine Pest hat, dann weiß er automatisch auch, dass die Freundin an Grippe erkrankt ist/war.
Um diese beiden Probleme anzugehen, wurde von Machanavajjhala et al. das Konzept der l-Diversity eingeführt. Die zugrundeliegende Idee ist, dass l-Diversity vorliegt, wenn es für jede Äquivalenzklasse l „wohl repräsentierte“ Werte für die sensiblen Daten gibt. Um nun wohl repräsentiert besser zu definieren gibt es drei Möglichkeiten l-Diversity im allgemeinen zu definieren. Betrachte man sich die einfachste Definition zuerst:

Definition 1.6 (Distinct l-Diversity [2])

In jeder Äquivalenzklasse gibt es mindestens l unterschiedliche Werte für die sensiblen Attribute.

Offensichtlich ist diese Definition ziemlich schwach und kann anhand mittels probabilistischer Inferenz angegriffen werden. Um also die statistische Verteilungen im Datensatz miteinzubeziehen kann l-Diversity mit Hilf von Informationsentropie definiert werden:

Definition 1.7 (Entropy l-Diversity [2])

Die Informationsentropy einer Äquivalenzklasse E ist definiert durch:

$$Entropy(E) = -\sum_{s \in S} p(E, s) \cdot log(p(E, s))$$

Dabei ist S die Menge der sensiblen Daten und $p(E, s)$ der Anteil der Elemente in E mit dem sensiblen Attribut s. Eine Tabelle hat Entropy l-Diversity, wenn $Entropy(E) \geq log(l)$ für jede Äquivalenzklasse E gilt.

Diese Definition setzt nun vorraus, dass die Informationsentropy der anonymisierten Tabelle bereits mindestens $log(l)$ ist. Diese Einschränkung kann nicht immer erfüllt werden. Daher gibt es noch folgende, weniger einschränkende Definition:

Definition 1.8 (Recursive (c,l)-Diversity [2])

Seien m Datensätze in einer Äquivalenzklasse E und sei r_i, $1 \leq i \leq m$ des i-häufigsten sensitiven Eintrages in E. E hat Recursive (c,l)-Diversity wenn $r_1 < c(r_l + r_{l+1} + ... + r_m)$. Gilt dies für alle Äquivalenzklassen, so hat die anonymisierte Tabelle Recursive (c,l)-Diversity.

Das ist ein praktisch verwendbarer Kompromiss mit welchem sichergestellt wird, dass unwahrscheinliche Werte nicht zu häufig in einer Äquivalenzklasse auftreten [2].

1.4 t-Closeness als Erweiterung von l-Diversity

l-Diversity ist zwar sicherer als k-Anonymität weißt aber für stark unausgeglichene Datensätze statistische Schwächen auf. Nehme man eine Anonymisierte Tabelle in der 99% negativ auf einen Virus getestet wurden. Erreichbar wäre hier 2-Diversity und eine Äquivalenzklasse mit 50% positiven und negativen sensiblen Daten wäre möglich. Weiß man, dass eine Person zu dieser Äquivalenzklasse gehört, so steigt die Zuversicht, dass derjenige positiv ist von 1% auf 50%.
Um derartige Angriffe auszuschließen wurde bei t-Closeness statt des Entropie-Ansatzes ein Wahrscheinlichkeitsverteilungs-Ansatz verwendet. Demnach kann t-Closeness im Allgemeinen so definiert werden:

Definition 1.9 (t-Closeness Grundprinzip [1])

Eine Äquivalenzklasse erfüllt t-Closeness, wenn der Unterschied zwischen der Verteilung der sensitiven Daten der Äquivalenzklasse zu der Verteilung der sensitiven Daten in der ganzen Tabelle kleiner als ein Schwellenwert t ist. Eine Tabelle erfüllt t-Closeness, wenn alle Äquivalenzklassen t-Closeness erfüllen.

Die Schwierigkeit bei t-Closeness ist die Berechnung des Unterschieds zwischen zwei Wahrscheinlichkeitsverteilungen. Dazu wurde sich für eine Berechnung basierend auf der Wasserstein-Metrik entschieden, welche im Englischen auch „Earth Mover Distance (EMD)“ bezeichnet wird [1].

1.5 Weiterentwicklungsmöglichkeiten von t-Closeness

Im folgenden werden Ansätze zur Weiterentwicklung von t-Closeness, welche im dazugehörigen Paper enthalten sind, kurz erläutert.
Mehrere sensible Attribute: Die Evaluierungsmöglichkeitn mittels t-Closeness ist für Tabellen mit mehreren sensiblen Attributen nicht eindeutig definiert. Hierzu könnten Ansätze entwickelt und analysiert werden [1].
Berücksichtigung von t-Closeness bei der Anonymisierung: Bei k-Anonymität und l-Diversity ist das Entfernen von Datensätzen meist wenig zielführen um eine optimal anonymisierte Tabelle zu erhalten. Bei t-Closeness kann dies aber nützlich sein, da sich hierbei für einen statistischen Ansatz entschieden wurde [1].
Grenzen der Wasserstein-Metrik: Die Wasserstein-Metrik ist zwar im Allgemeinen eine gute Möglichkeit um die Unterschiede zwischen zwei Wahrscheinlichkeitverteilungen zu berechnen, aber diese ist dennoch nicht perfekt. Es kann demnach noch analysiert werden ob es im Allgemeinen oder für ausgewählte Spezialfälle bessere Möglichkeiten gibt [1].

1.6 Problemschwierigkeit von k-Anonymität und Erweiterungen

Mittels Generalisierung kann k-Anonymität erreicht werden. Dadurch entsteht eine sehr hohe Variabilität in der Menge an möglichen anonymisierten Ergebnissen, welche unterschiedlich optimal sind. Das optimale Ergebnis zu finden erweist sich dabei als äußerst schwierig. Es konnte nämlich durch Reduktion auf das k-Dimensionale Matching Problem bewiesen werden, dass die Realisierung von k-Anonymität mit $k > 2$ ein NP-schweres Problem ist [3].

Literatur

[1] Li, Ninghui ; Li, Tiancheng ; Venkatasubramanian, Suresh: t-closeness: Privacy beyond k-anonymity and l-diversity. In: *2007 IEEE 23rd International Conference on Data Engineering* IEEE, 2007, S. 106–115

[2] Machanavajjhala, Ashwin ; Kifer, Daniel ; Gehrke, Johannes ; Venkitasubramaniam, Muthuramakrishnan: l-diversity: Privacy beyond k-anonymity. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1 (2007), Nr. 1, S. 3–es

[3] Meyerson, Adam ; Williams, Richard: General k-Anonymization is Hard. (2003), 04

[4] Sweeney, Latanya: k-anonymity: A model for protecting privacy. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (2002), Nr. 05, S. 557–570