Differential Privacy bei korrelierten Daten

Herausforderungen und das Konzept Pufferfish

Rafael Vrečar
Technische Universität Wien
Matrikelnummer: 01627765
Kontakt: e01627765@student.tuwien.ac.at

19. Juli 2018

Abstract
Differential Privacy ist ein Ansatz, um zu gewährleisten, dass Daten mit einem gewissen, messbaren Level an Privatsphäre geschützt sind. Dabei wird den Daten sogenanntes Rauschen beigefügt, um Personen, zu denen ein Datensatz gehört, nicht mehr eindeutig identifizierbar zu machen. Bei Daten, die von einander unabhängig sind, ist dieses Konzept zielführend, doch wenn Daten miteinander korrelieren, reicht derartiges Rauschen nicht aus, da aufgrund der Korrelationen Rückschlüsse auf die zugehörigen Personen gezogen werden können. Ebenso ist es problematisch, wenn die Daten von einer einzelnen Person mit Rauschen versehen werden sollen. Ein generalisiertes Konzept der Differential Privacy ist Pufferfish. Dieses Framework verspricht, Abhilfe für die angesprochenen Probleme zu schaffen.

1 Einführung

Durch die Digitale Revolution ist der Mensch mit einer Unmenge an Daten konfrontiert. Sobald man im Internet nach einem Begriff sucht, mit der Bankomatkarte einen Einkauf tätigt oder per App eine Zugfahrkarte löst, werden Daten erhoben und gespeichert. Diese Daten stellen für Unternehmen ein unfassbar wertvolles Gut dar, da sie vieles über ihre Kundinnen und Kunden aussagen. Diese Daten können dann beispielsweise für personalisierte Werbung oder persönliche Sonderangebote genutzt werden, aber auch einfach, um das Angebot für gewisse Zielgruppen zu optimieren.
Medizinische Einrichtungen können Daten verwenden, um die Behandlungsmöglichkeiten zu verbessern, gewisse Krankheitsbilder zu beforschen und Zusammenhänge zwischen verschiedenen Krankheiten, Lebensgewohnheiten etc. herzustellen.

Wissenschaft und Wirtschaft leben also von den Daten, die durch die Digitale Revolution leichter erhebbar, speicherbar und abrufbar geworden sind.

Doch gerade bei sensiblen Informationen wie den angesprochenen Daten zu Krankheiten etc. ist es notwendig, Vorsicht walten zu lassen, um die Integrität der Privatsphäre eines Menschen zu wahren. Eine gängige Variante hierbei ist es, die Daten zu verschlüsseln. Dabei werden die Daten über komplexe mathematische Funktionen in eine Form gebracht, die für den Menschen keine Aussagekraft hat. Mithilfe eines sogenannten Schlüssels können die Daten wieder in eine lesbare Form gebracht werden. Die Kenntnis des Schlüssels ist also, in einem optimalen Verschlüsselungssystem, Voraussetzung, um die Daten lesen zu können.
Nun sind die Daten zwar vor Angreiferinnen und Angreifern geschützt, aber bei der Verarbeitung ist es notwendig, sie unverschlüsselt zu verwenden, da sie sonst bedeutungslos sind. Sensible Daten würden also den Forscherinnen und Forschern etc. offen liegen und damit würde die Privatsphäre verletzt werden. Ein Ansatz, der dieses Problem behebt, ist *Differential Privacy*.

Diese Arbeit beschäftigt sich mit besagtem Ansatz. Nach einer generellen Einführung in das Thema wird der Hintergrund zu Differential Privacy beispielhaft erläutert. Es wird außerdem erklärt, was korrelierte Daten sind. Danach wird die Differential Privacy formalisiert und es werden die Konzepte ϵ-Differential Privacy und (ϵ, δ)-Differential Privacy sowie deren Probleme bei korrelierten Daten erläutert. In weiterer Folge wird darauf aufbauend das Konzept Pufferfish erklärt, ehe mit einer Konklusion geschlossen wird.

2 Hintergrund zu Differential Privacy

Differential Privacy ist ein Ansatz zur Wahrung der Privatsphäre von Daten, ohne diese zu verschlüsseln. Hierbei wird den Daten ein sogenanntes Rauschen, also eine bewusste Veränderung, beigefügt. Dadurch ist unmöglich feststellbar, ob ein Datensatz zu einem gewissen Individuum gehört. Wir werden später sehen, dass es steuerbar ist, wie viel Rauschen beigefügt wird und wie stark die Daten dadurch verändert werden [**7**; *#1. INTRODUCTION, 1. Absatz*].
Ziel dabei ist es, beispielhaft gesprochen, viel über eine Population als Ganzes zu eruieren, ohne dabei viel über ein einzelnes Individuum der Population zu erfahren [**3**; *#1 The Promise of Differential Privacy, 2. Absatz*].

Wir sehen uns zur eben getätigten Aussage ein Beispiel an. Angenommen, wir besitzen eine Datenbank, in der wir die politische Präferenz für die nächste Nationalratswahl von 100.000 wahlberechtigten Österreicherinnen und Österreichern gespeichert haben. Wir haben dabei den vollständigen Namen, das Geschlecht, das exakte Geburtsdatum, die genaue Wohnadresse und exakte Tätigkeit mit beispielsweise Arbeitgeber, Position etc. gespeichert. Um relevante Aussa-

gen über das Wahlverhalten zu treffen, ist es unerheblich, wie die Person heißt, wann genau sie geboren ist, wo genau sie wohnt und was genau sie beruflich macht. Wenn wir nun den Namen aus unseren Datensätzen entfernen, das Geburtsdatum durch das Alter bzw. ein Altersintervall ersetzen, die Wohnadresse durch den Bezirk und die genaue Tätigkeit durch eine Berufs- bzw. Tätigkeitsgruppe, dann können wir immer noch Aussagen wie beispielsweise *„Studierende im Alter von 20 bis 25 Jahren, die im Bezirk Wieden leben, wählen zu 25 Prozent Partei A.“* treffen, ohne, dass eine Person auch nur im Ansatz eindeutig identifiziert werden könnte.
Diese einfache Modifikation ist deshalb möglich und statistisch gesehen bedeutungslos, da die Datensätze von einander unabhängig sind. Wenn wir den Datensatz von Person A verändern oder löschen, dann beeinflusst das den Inhalt des Datensatzes von Person B nicht.

Korrelierte Daten

Schwieriger ist das bei Datensätzen, die miteinander verknüpft sind, sprich miteinander korrelieren. Ein Beispiel hierfür sind befreundete Personen in sozialen Netzwerken [**4**; *#1. INTRODUCTION, 9. Absatz*]. Angenommen, Person A ist mit den Personen B und C befreundet, die ebenfalls miteinander befreundet sind. So können, selbst, wenn sämtliche Daten von Person A gelöscht werden, und nur noch gespeichert bleibt, dass Person B und C jeweils Person A kennen, Rückschlüsse auf Person A gezogen werden, wenn man B und C genauer kennt.

Ein weiteres Beispiel dieser Art ist das Aufzeichnen der Aufenthaltsorte eines Menschen. Man kann dann beispielsweise Statistiken führen, wie viele Menschen sich gerade gleichzeitig an einem Ort aufhalten. Es sei zusätzlich realitätsgetreu vorausgesetzt, dass sich eine Person zu einem Zeitpunkt nur an einem Ort aufhalten kann. Selbst beim Hinzufügen von Rauschen können Rückschlüsse gezogen werden, wo sich ein Mensch gerade aufhält. Hilfreich hierbei sind Straßennetzwerke und die Kenntnis der Gewohnheiten einer Person. Wenn bekannt ist, dass eine Person Ort Y immer nach Ort X besucht, zum Beispiel auf dem Arbeitsweg, und man zusätzlich weiß, wie weit die beiden Orte voneinander entfernt sind, können trotz Veränderung der Daten durchaus Prognosen angestellt werden [**1**; *#I. INTRODUCTION, Example 1*].
Die eben angesprochene Problematik lässt sich durch eine einfache Modifikation nur schwer umschiffen. Wir werden uns später ein Konzept ansehen, das diese Fälle besser abdeckt als die gängigen Konzepte der Differential Privacy, doch zunächst werden wir zwei grundlegende Konzepte der Differential Privacy motivieren.

3 Formalisierung der Differential Privacy

Im Folgenden werden wir zunächst den Begriff Differential Privacy formalisieren und uns dann zwei grundlegende Konzepte der Differential Privacy genauer ansehen, ihren Aufbau analysieren, ihre Stärken und Schwächen betrachten und dazu überleiten, warum diese Konzepte bei korrelierten Daten an ihre Grenzen stoßen. Differential Privacy bietet Privatsphäre durch die Einführung von Zufall. Dies lässt sich mit dem folgenden Beispiel illustrieren: Studierende werden dazu angehalten, auszusagen, ob sie eine Eigenschaft P besitzen oder nicht und zwar nach folgenden Anweisungen:

1. *Wirf eine Münze.*
2. *Liegt „Zahl“ oben, dann antworte wahrheitsgetreu.*
3. *Liegt „Kopf“ oben, dann wirf eine weitere Münze und antworte „Ja“, falls „Kopf“ oben liegt und „Nein“, falls „Zahl“ oben liegt.*

Weil nicht sicher ist, ob eine Aussage der Wahrheit entspricht, wird Privatsphäre erzeugt. Beispielsweise tritt die Antwort „Ja“ mit einer Häufigkeit von $\frac{1}{4}$ auf, obwohl ein/e Studierende/r die besagte Eigenschaft nicht hat. Anschaulich gesprochen beträgt die Wahrscheinlichkeit, dass die Antwort der Wahrheit entspricht also 75 Prozent [**3**; *#2.3 Formalizing differential privacy, 1. Absatz ff.*]. Man kann diesen Ablauf als Funktion beschreiben. Die Grundmenge sind dabei die Studierenden, die Werte- bzw. Zielmenge lautet $\{true, false\}$. Die Funktion erhält die Eingabe *Studierende/r X* und gibt aus, ob diese/r über die Eigenschaft P verfügt. Das Besondere an dieser Funktion ist, dass sie randomisiert ist. Das heißt, das Ergebnis wird nicht nur von der Eingabe, sondern auch vom Zufall beeinflusst. In diesem Fall ist das der Münzwurf. Man kann hier natürlich ebenso Zufallsgeneratoren oder andere Mittel bemühen, was in der theoretischen Betrachtung aber irrelevant ist. Wir nennen die randomisierte Funktion im Folgenden f.
Genauere Definitionen sind der eben genannten Quelle zu entnehmen. In unserem Fall genügt aber die oben angeführte Beschreibung. Darauf aufbauend betrachten wir nun das erste Konzept der Differential Privacy, die *ϵ-Differential Privacy*.

ϵ-Differential Privacy

Wir beginnen direkt mit der Definition, bevor sie stückweise erläutert wird [**2**; *#Differential Privacy, Definition 2*].

Definition: *Besagte Funktion f liefert ϵ-Differential Privacy, wenn für zwei Datensätze D_1 und D_2, die sich in höchstens einem Element (z. B. eine Zeile in einer relationalen Datenbank) unterscheiden, und alle $S \subseteq W$ (S ... Bildmenge von f, W ... Wertemenge von f) gilt:* $Pr[f(D_1) \in S] \leq exp(\epsilon) \times Pr[f(D_2) \in S]$

Diese Definition bedeutet, dass ϵ-Differential Privacy dann erfüllt ist, wenn die Wahrscheinlichkeit, dass das Ergebnis von f angewendet auf den Datensatz D_1 mit einer Wahrscheinlichkeit in S liegt, die kleiner oder gleich e^{ϵ} mal der Wahrscheinlichkeit, dass das Ergebnis von f angewendet auf den Datensatz D_2 in S liegt.

Die bereits angesprochene Funktion f, die diese Definition erfüllt, stellt, anschaulich gesprochen, Folgendes sicher:
Person A möchte sich bei einer Versicherungsgesellschaft versichern lassen. Die Versicherungsgesellschaft verfügt über eine Datenbank an sensiblen Informationen, die mit der oben definierten ϵ-Differential Privacy gesichert ist. In diesem Fall macht es für Person A keinen signifikanten Unterschied für den Erhalt einer Versicherung, ob sie in der Datenbank enthalten ist oder nicht [**2**; *#Differential Privacy, Definition 2 ff.*].

Wenn wir den Datenbestand sowie die Funktion f als gegeben annehmen, dann kann einzig und allein durch den Parameter ϵ die „Stärke“ der Privatsphäre variiert werden. Dieser ist aber auch konstant, sofern man zusichern will, dass der Datenbestand mit ϵ-Differential Privacy geschützt ist. Es zeigt sich, dass dieser Schutz Daten für gewisse Anwendungsfälle unbrauchbar machen kann, da die Privatsphäre zu stark geschützt ist. Dieses Problem lässt sich durch eine Erweiterung, die sogenannte (ϵ, δ)-*Differential Privacy*, umschiffen.

(ϵ, δ)-Differential Privacy

Der Parameter δ wird dabei zur rechten Seite der Ungleichung addiert. Das hat folgende leicht veränderte Definition zur Folge [**3**; *#2.3 Formalizing differential privacy, Definition 2.4 ff.*].

Definition: *Die bereits angesprochene Funktion f liefert (ϵ, δ)-Differential Privacy, wenn für zwei Datensätze D_1 und D_2, die sich in höchstens einem Element (z. B. eine Zeile in einer relationalen Datenbank) unterscheiden, und alle $S \subseteq W$ (S ... Bildmenge von f, W ... Wertemenge von f) gilt:*
$Pr[f(D_1) \in S] \leq exp(\epsilon) \times Pr[f(D_2) \in S] + \delta$

δ lässt nun also zu, dass die ursprünglichen Bedingungen bis zu einem gewissen Grad verletzt werden dürfen. Wie sehr, hängt von der Größe von δ ab. Wie aus der Ungleichung hervorgeht, sind die erlaubten Abweichungen desto größer, je größer δ ist.

Probleme dieser Konzepte bei korrelierten Daten

Grundsätzlich lässt sich feststellen, dass Differential Privacy in dieser Form für viele Anwendungsfälle zielführend ist. Dennoch stoßen wir bei korrelierten Daten auf Probleme. Dazu folgender Gedankengang:
Es soll die körperliche Aktivität einer Person über einen längeren Zeitraum aufgezeichnet werden und daraus aggregierende Statistiken erstellt werden. Gleichzeitig soll aber verborgen bleiben, zu welchem exakten Zeitpunkt eine Aktivität stattgefunden hat. Wenn die Messung in kurzen Zeitintervallen durchgeführt wird, dann führt das zum Problem, dass stark korrelierte Zeitreihen auftreten, da sich die menschlichen Aktivitäten vergleichsweise langsam ändern. Ähnlich wie das oben angesprochene Beispiel mit dem Aufzeichnen des Ortes, an dem sich ein Mensch befindet, wird man beispielsweise nicht in einer Millisekunde einen Berg besteigen und in der nächsten ein heißes Bad nehmen. Da es sich hierbei um die Daten eines einzelnen Menschen handelt, ist Differential Privacy nicht wie oben beschrieben anwendbar [**7**; *#1. INTRODUCTION, 2. Absatz*].
Es gibt hierfür zwar weitere Ansätze, die sogenannte *entry-privacy* und die sogenannte *group differential privacy*. Diese werden in dieser Arbeit aber nicht weiter behandelt. Sie werden in der unten angeführten Quelle angesprochen [**7**; *#1. INTRODUCTION, 3. Absatz*].

Wir wollen uns jetzt einem etwas generellerem Konzept der Differential Privacy widmen, welches Probleme bei korrelierten Daten behebt. Dieses Konzept heißt *Pufferfish*.

4 Pufferfish

Pufferfish ist eine generalisierte Version der Differential Privacy. Das Framework und wurde von D. Kifer und A. Machanavajjhala vorgestellt und kann verwendet werden, um spezifische Definitionen betreffend die Privatsphäre für gewisse Anwendungsfälle zu erstellen. Ziel dieses Frameworks sei, Expertinnen und Experten ohne fundiertes Wissen bezüglich Privatsphäre, auf Anwendungsebene zu ermöglichen, starke Definitionen zu erstellen. Zusätzlich kann Pufferfish auch dazu verwendet werden, bereits bestehende Definitionen zu untersuchen [**5**; *1. Absatz*].

Definition: *Bei Pufferfish sind die Voraussetzungen durch drei Komponenten spezifiziert. S ist eine Menge an sogenannten Geheimnissen. S repräsentiert, was geschützt werden muss. Q ist eine Menge an geheimen Paaren. Q repräsentiert Paare von Geheimnissen, die nach außen ununterscheidbar sein müssen. Θ ist eine Klasse von Distributionen für die Daten-Generierung* [**7**; *#1. INTRODUCTION, 4. Absatz*].

Privatsphäre wird dadurch erreicht, dass sichergestellt wird, dass die geheimen Paare in Q ununterscheidbar sind, wenn Daten aus jedem beliebigen $\theta \in \Theta$ generiert werden.

Im oben genannten Zeitreihen-Beispiel, wo die Aktivität eines Menschen über einen längeren Zeitraum aufgezeichnet werden soll, ist S die Menge an Aktivitäten zu jedem Zeitpunkt t und die geheimen Paare sind Tupel der Form *(Aktivität a zum Zeitpunkt t, Aktivität b zum Zeitpunkt t)*. Θ ist dabei eine Menge an sogenannten Markov-Ketten. Diese sind eine speziell Form von Zufallsprozessen. Zufallsprozesse sind die mathematischen Beschreibungen von zufälligen Vorgängen, die aber zeitlich geordnet sind. Sie werden in dieser Arbeit jedoch nicht weiter erläutert [**7**; *#1. INTRODUCTION, 4. Absatz*].

Pufferfish erfasst Korrelationen in Anwendungsfällen wie dem oben angesprochenen auf zwei Arten. Erstens kann Pufferfish im Gegensatz zur Differential Privacy private Werte der Datensätze vor Korrelationen mit verschiedenen Einträgen bzw. Individuen schützen. Zweitens lässt es sich auch in Anwendungsfällen verwenden, wo eine große Anzahl an Datensätzen korrelieren, wobei die durchschnittliche Menge an Korrelation aber gering ist [**7**; *#1. INTRODUCTION, 4. Absatz*]. Die größte Herausforderung bei der Verwendung von Pufferfish ist, dass es nicht viele passende *Mechanismen* gibt. Ein Mechanismus ist in diesem Zusammenhang ein Algorithmus, dessen Output es Anwenderinnen und Anwendern erlaubt, statistische Daten zu analysieren. Das Verhalten eines solchen Algorithmus ist über sogenannte Privatsphäre-Definitionen festgelegt.
Es gibt zwar gewisse Mechanismen für spezielle Instanzen, aber für generelle Algorithmen gibt es bis dato nur eine relativ geringe Zahl an Ansätzen. In der unten genannten Quelle wird der sogenannte *Wasserstein Mechanismus* vorgestellt. Aufgrund von Ineffizienz wird in diesem Zusammenhang auch der Fall besprochen, in welchem Korrelationen zwischen Variablen durch ein sogenanntes *Bayessches Netz* beschrieben werden können, um die Komplexität mithilfe eines weiteren Mechanismus, dem sogenannten *Markov Quilt Mechanismus*, zu reduzieren. Eine detaillierte Behandlung der gerade eben erwähnten Algorithmen würde den Rahmen dieser Arbeit sprengen, insofern bleibt es an dieser Stelle bei der Benennung [**7**; *#1. INTRODUCTION, 6. Absatz* & **5**; *#1. INTRODUCTION, 1. Absatz*].

Beispiele für Anwendungsfälle von Pufferfish

Im Folgenden werden wir zwei einfache Beispiele, die aber die Anwendbarkeit von Pufferfish, sehr anschaulich erläutern, betrachten.

Beispiel 1: Überwachung der körperlichen Aktivität einer Person

Dieses Beispiel haben wir bereits vorher angesprochen und festgestellt, dass Differential Privacy hier nicht direkt anwendbar ist. In diesem Zusammenhang besteht die Datenbank aus einer Zeitreihe $X = \{X_1, X_2, ..., X_T\}$, wobei ein Eintrag X_t die physische Aktivität, wie zum Beispiel Laufen, Sitzen etc., zum Zeitpunkt t beschreibt. Das Ziel dabei ist, ein approximiertes

Histogramm für einen gewissen Zeitraum, beispielsweise eine Woche, zu veröffentlichen, ohne dabei die konkrete Aktivität der Person zu einem konkreten Zeitpunkt preiszugeben [**7**; *#2. THE SETTING, Example 1*]. Wir definieren die Menge S als $\{s_a^t : t = 1, ..., T, a \in A\}$, wobei gilt, dass A die Menge aller Aktivitäten ist und $1, ..., T$ die Zeitpunkte darstellt. Die Aktivität zum Zeitpunkt t ist also ein Geheimnis. Die Menge Q besteht hierbei aus Paaren (s_a^t, s_b^t), wobei gilt $a, b \in A$ und t ein Zeitpunkt ist. Das bedeutet, dass nicht festgestellt werden kann, ob eine Person zum Zeitpunkt t nun Aktivität a oder Aktivität b ausführt für alle Paare dieser Art. Θ ist in diesem Beispiel eine Menge an Zeitreihen-Modellen, die beschreiben, wie die Personen zwischen den Aktivitäten wechseln. Plausibel ist in diesem Zusammenhang zum Beispiel die Modellierung mit einer Menge aus Markov-Ketten $X_1 \to X_2 \to ... \to X_T$, wobei jeder Zustand X_t eine Aktivität in A ist. Das Konzept der Markov-Ketten wird in dieser Arbeit aber, wie oben bereits erwähnt, nicht näher beleuchtet.
Differential Privacy wäre in diesem Beispiel deshalb nicht zweckmäßig gewesen, da es sich um die Werte einer einzigen Person handelt [**7**; *#2. THE SETTING, 2.2 Examples, Example 1*].

Beispiel 2: Status der Grippeerkrankung einer Person

Hier besteht die Datenbank aus einer Menge an Wahrheitswerten $X = \{X_1, X_2, ..., X_n\}$. Dabei sagt X_i aus, ob Person i krank ($X_i = 1$) oder gesund ($X_i = 0$) ist. Das Ziel ist nun, eine Approximation zur Summe $\sum_{i=0}^{n}$ der infizierten Personen zu veröffentlichen, während jedoch sichergestellt ist, dass auf Basis dieser Veröffentlichung nicht herausgefunden werden kann, ob die spezielle Person i infiziert ist oder nicht. Weiter ist festzuhalten, dass die Datenbank Personen beinhaltet, die miteinander interagieren, also zum Beispiel den gleichen Arbeitsplatz haben, befreundet sind etc. und sich somit regelmäßig sehen. Daher korrelieren die Status der Grippeerkrankungen verschiedener Personen stark. Zusätzlich wird die Entscheidung, bei der Datenerhebung mitzumachen, gruppenweise getroffen (zum Beispiel vom Arbeitgeber, in der Schule etc.) und daher kann eine einzelne Person nicht entscheiden, ob sie mitmacht oder nicht [**7**; *#2. THE SETTING, Example 2*]. In diesem Beispiel verwenden wir eine etwas andere Notation. s_0^i bedeutet hierbei, dass Person i keine Grippe hat, während s_1^i bedeutet, dass sie die Grippe hat. Θ ist in diesem Zusammenhang eine Menge von Modellen, welche die Verbreitung der Krankheit beschreiben. Ein mögliches Element θ dieser Menge ist ein Tupel (G_θ, p_θ), wobei $G_\theta = (X, E)$ ein Graph zwischenmenschlicher Interaktionen ist und p_θ eine Wahrscheinlichkeitsverteilung der Komponenten von G_θ darstellt. Als konkretes Beispiel könnte man G_θ als eine Vereinigung von Freundeskreisen $F_1, F_2, ...F_N$ sehen, während p_θ eine konkrete Wahrscheinlichkeitsverteilung von der Anzahl der infizierten Personen in jedem Freundeskreis ist.
In diesem Fall würde der klassische Ansatz von Differential Privacy zwar genug Rauschen hinzufügen, um zu verbergen, ob eine konkrete Person im Datensatz präsent ist, aber über die Korrelation der Datensätze könnten trotzdem Rückschlüsse gezogen werden, ob besagte Person infiziert ist oder nicht [**7**; *#2. THE SETTING, 2.2 Examples, Example 2*].

5 Konklusion

Zusammenfassend ist zu sagen, dass die Privatsphäre in der besprochenen Ausprägung dadurch essenziell wurde, dass wir durch die Digitale Revolution die Möglichkeit haben, unzählige Daten effizient zu speichern und zu verarbeiten.

Um ein zufriedenstellendes Level an Privatsphäre zu gewährleisten, ist die Wahl eines passenden Ansatzes entscheidend. Ob ein Ansatz passend ist, hängt in erster Linie davon ab, Daten welcher Art gespeichert sind, wie diese zusammenhängen und wie hoch der Schutz dieser Daten sein soll.

In diesem Zusammenhang ist festzustellen, dass der klassische Ansatz der Differential Privacy, also die ϵ-Differential Privacy sowie die Erweiterung namens (ϵ, δ)-Differential Privacy, in vielen Anwendungsfällen sehr gute Ergebnisse liefert. Beispielsweise liefert der klassische Ansatz bei der Anwendung auf voneinander unabhängigen Datensätzen, die nicht miteinander korrelieren, sehr gute Ergebnisse. Zu Problemen kommt es allerdings, wenn die Daten miteinander korrelieren bzw. die Daten nur von einem einzigen Individuum stammen. Diese Probleme werden vom Privatsphäre-Framework Pufferfish umgangen. Dabei ist in erster Linie auf die Wahl eines passenden Mechanismus Acht zu geben.

Abschließend ist noch anzumerken, dass Privatsphäre ein hohes Gut für jeden Menschen sein sollte und dementsprechend auch die Forschung in diesem Gebiet höchst relevant ist.

Weitere interessante Ansätze und Konzepte sind in den Quellen [**6** & **8**] zu begutachten.

Literaturverzeichnis

Zitiert wurde jeweils im von der Quelle bereitgestellten Stil (siehe *https://dl.acm.org/* bzw. *http://dblp.uni-trier.de*). Zuletzt abgerufen wurden die beiden Seiten jeweils am 12.05.2018.

Sofern die PDF-Dokumente nicht über *The ACM Digital Library (https://dl.acm.org/)* abgerufen wurden, ist der entsprechende Link zusätzlich angegeben.
Alle Links, bei denen das Datum des letzten Abrufs nicht genauer spezifiziert ist, wurden zuletzt am 10.06.2018 abgerufen.
Die Links wurden, sofern möglich, mit *https* angegeben.

[1] Yang Cao, Masatoshi Yoshikawa, Yonghui Xiao and Li Xiong. Quantifying Differential Privacy under Temporal Correlations. In 33rd IEEE International Conference on Data Engineering, ICDE 2017, San Diego, CA, USA, April 19-22, 2017. IEEE Computer Society, 821-832.
DOI: https://dblp.org/rec/bib/conf/icde/CaoYX017
[LINK: https://ieeexplore.ieee.org/document/7930028/,
zuletzt abgerufen am 10.06.2018]

[2] Cynthia Dwork. 2006. Differential privacy. In Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II (ICALP'06), Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg, 1-12.
DOI=https://dx.doi.org/10.1007/11787006_1
[LINK: https://link.springer.com/content/pdf/10.1007/11787006.pdf,
zuletzt abgerufen am 10.06.2018]

[3] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. Found. Trends Theor. Comput. Sci. 9, 3–4 (August 2014), 211-407.
DOI=https://dx.doi.org/10.1561/0400000042
[LINK: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf,
zuletzt abgerufen am 12.05.2018]

[4] Daniel Kifer and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD '11). ACM, New York, NY, USA, 193-204.
DOI: https://doi.org/10.1145/1989323.1989345

[5] Daniel Kifer and Ashwin Machanavajjhala. 2014. Pufferfish: A framework for mathematical privacy definitions. ACM Trans. Database Syst. 39, 1, Article 3 (January 2014), 36 pages.
DOI=https://dx.doi.org/10.1145/2514689

[6] André L. C. Mendonça, Felipe T. Brito, Leonardo S. Linhares, and Javam C. Machado. 2017. DiPCoDing: A Differentially Private Approach for Correlated Data with Clustering. In Proceedings of the 21st International Database Engineering & Applications Symposium (IDEAS 2017), Bipin C. Desai, Jun Hong, and Richard McClatchey (Eds.). ACM, New York, NY, USA, 291-297.
DOI: https://doi.org/10.1145/3105831.3105861

[7] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. 2017. Pufferfish Privacy Mechanisms for Correlated Data. In Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17). ACM, New York, NY, USA, 1291-1306.
DOI: https://doi.org/10.1145/3035918.3064025

[8] Sen Su, Peng Tang, Xiang Cheng, Rui Chen and Zequn Wu. 2016. Differentially private multi-party high-dimensional data publishing. In 32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016. IEEE Computer Society, 205-216.
DOI: https://doi.org/10.1109/ICDE.2016.7498241
[LINK: https://ieeexplore.ieee.org/document/7498241/,
zuletzt abgerufen am 10.06.2018]

Acknowledgement

Das Template dieser Arbeit ist von *https://www.latextemplates.com/template/science-journal* und unter der *CC Attribution-NC-SA-Lizenz* [*https://creativecommons.org/licenses/by-nc-sa/3.0/*] veröffentlicht worden.
Dabei wurden kleinere Änderungen, wie der Zeilenabstand und die Formatierung beim Abstract, vorgenommen und nicht benötigte Passagen entfernt. Die oben angeführten Links wurden jeweils zuletzt am 12.05.2018 abgerufen.

Erklärung zum Verfassen dieser Arbeit

Hiermit erkläre ich, Rafael Vrečar, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und, dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe."
Diese Erklärung (Vorlage der Technischen Universität Wien) wurde elektronisch per E-Mail am 19. Juni 2018 aus Wien abgegeben und ist ohne Unterschrift gültig.