VERFÜGBARKEIT VON PERSONENBEZOGENEN DATEN UNTER DEM ASPEKT DER ANONYMISIERUNG

Fabian Engl
Ostbayerische Technische Hochschule Regensburg
Galgenbergstraße 32, 93053 Regensburg
Email: fabian1.engl@st.oth-regensburg.de

***Abstract*—Daten und Analyseverfahren bestimmen zunehmend den heutigen beruflichen sowie privaten Alltag. In den letzten Jahren gewannen besonders personenbezogene Daten immer mehr Relevanz. Während hyper-personalisierte Werbung auf Social Media immer wieder in die Kritik gerät, profitieren auch andere Bereiche von leicht zugänglichen persönlichen Informationen. Die medizinische Forschung oder die Entwicklung künstlicher Intelligenzen hängen enorm von der Verfügbarkeit derartiger Daten ab. Neue Gesetzgebungen und Normen schränken diese aber immer weiter ein und verpflichten die Anonymisierung von personenbezogenen Daten vor deren Verwendung.**

In diesem Artikel erfolgen eine Auflistung aktueller Anonymisierungsverfahren sowie eine Analyse ihrer Schwachstellen. Letztere werden kritisch unter dem Risiko der Re-Identifizierung und dem einhergehenden Verlust an Datenqualität beurteilt. Die Untersuchung veranschaulicht, dass alle Verfahren die Aussagekraft von personenbezogenen Daten auf unterschiedliche Art und Weise vermindern. Der Artikel identifiziert dabei drei wichtige Anforderungskriterien an einen anonymen Datensatz und hebt die Bedeutung einer ausführlichen Auseinandersetzung mit diesen Kriterien hervor. So bedarf es möglicherweise der Kombination mehrerer Verfahren, um die Verfügbarkeit zu optimieren. Er schließt mit der These, dass unterschiedliche personenbezogene Attributklassen eine Auswirkung auf die Wahl eines passenden Anonymisierungsverfahrens haben können.

I. Problembeschreibung

Mit der Einführung neuer Digitalgesetze wie etwa der Datenschutz-Grundverordnung (DSGVO) nimmt der Schutz persönlicher Daten immer wieder eine zentrale Stellung in der Informationssicherheit ein. So garantiert die DSGVO Verbrauchern nicht nur die Aufrechterhaltung ihrer Privatsphäre, sondern verpflichtet Dritte sämtliche Verarbeitungsprozesse von personenbezogenen Daten offenzulegen (Bieker, Bremert und Hansen 2018). Als direkte Folge daraus mussten viele Betriebe ihre internen Datenauswertungen neu evaluieren und entsprechende Maßnahmen ergreifen.

Um dieses Problem zu lösen und den Schutz persönlicher Daten in allen Verarbeitungsschritten zu gewährleisten, kommen Anonymisierungsverfahren zum Einsatz. Deren zunehmende Bedeutung ist besonders an den Google Trends erkenntlich. Hier kam es im Mai 2018 nach Inkrafttreten der DSGVO zu einem starken Anstieg der Suchanfragen zum Thema *Anonymisierung und Pseudonymisierung*. Das Nachfrageniveau verdoppelte sich im Vergleich zum Zeitraum vor der Einführung (Google LLC 2022).

Die zunehmende Relevanz von Daten in allen Lebensbereichen erfordert einen stärkeren Fokus auf den Schutz von Individuen. Folglich ist eine kritische Auseinandersetzung mit existierenden Anonymisierungsansätzen sowie deren Auswirkungen auf die Verfügbarkeit von persönlichen Informationen notwendig. Dieser Artikel gibt zunächst einen Überblick über aktuelle Anonymisierungsverfahren, geht kurz auf deren Funktionsweise sowie Schwächen ein und zieht anschließend ein Fazit, wie stark deren Anwendung die Datenverfügbarkeit beeinflusst.

II. Funktionsweise und Schwächen existierender Anonymisierungsverfahren

Derzeit existieren viele verschiedene Methoden, um personenbezogene Daten zu anonymisieren. Alle Ansätze versuchen, den Bezug zwischen Personen und weiteren zu ihnen gehörenden Informationen unkenntlich zu machen. Die Trennung ist erreicht, wenn eine Person nicht mehr eindeutig durch einen Datensatz identifiziert werden kann. Bei Menschen unterscheidet man zwischen direkten und indirekten Identifikatoren (Kesarwani u. a. 2021). Ein Beispiel für Erstere sind Namen und private Kennnummern. Da diese ein Individuum eindeutig bestimmen, müssen sie vor der weiteren Verwendung entfernt werden. Indirekte Identifikatoren – auch Quasi-Identifikatoren (QI) genannt – lassen durch ihre Kombination Rückschlüsse auf eine einzelne Person oder eine kleine Personengruppe zu. Dazu zählen etwa das Geburtsdatum oder Postleitzahlen. Des Weiteren existieren sensible Daten wie beispielsweise eine Krankheit, die einer Person zuordenbar sind. Bei der Unkenntlichmachung sollen dabei möglichst viele Merkmale sowie deren Aussagekraft beibehalten werden.

A. *k*-Anonymität

Der Ansatz der k-Anonymität schreibt vor, dass jede Kombination der QI mindestens k-mal im Datensatz existieren muss (Kesarwani u. a. 2021). Auf diese Weise kann ein Angreifer selbst nach Reduzierung des Datensatzes die korrekte Person höchstens mit einer Wahrscheinlichkeit von $1/k$ identifizieren. Je höher der Wert k, desto schwieriger fällt die korrekte Bestimmung.

Dieser Ansatz weist allerdings viele Schwächen auf: Entsprechen ein oder mehrere Einträge nicht der k-Restriktion, müssen diese vollständig weggelassen oder deren Werte generalisiert werden. Beides verringert die Aussagekraft und verschlechtert die Datenqualität (Dubli

und Yadav 2017). Extreme Merkmalsausprägungen erschweren zudem die Definition von sinnvollen Wertebereichen (Li, Qardaji und Su 2011). Problematisch sind ebenfalls *k*-anonyme Gruppen, bei denen alle Mitglieder das gleiche sensible Merkmal wie etwa identische Diagnosen aufweisen (Krebs und Hagenweiler 2022). Bei dynamischen Zugriffen bietet sie zudem keinen Schutz vor *Brute-Force*-Angriffen, die gezielt QI-Kombinationen abfragen, um so Gruppen weiter aufzubrechen (Li, Qardaji und Su 2011). Das vorausschauende Erkennen von Abhängigkeiten zwischen Abfragen und deren systemseitiger Ablehnung, erfordert mit zunehmender Attributvielfalt enorme Rechenkapazitäten (Kesarwani u. a. 2021).

Aus diesen Gründen gilt die *k*-Anonymität alleine als zu schwache Form der Anonymisierung und benötigt in der Praxis eine Ergänzung durch andere Verfahren (Li, Qardaji und Su 2011).

B. ℓ-Diversity and t-Closeness

Der *ℓ-Diversity*-Ansatz erweitert die *k*-Anonymität durch eine zusätzliche Einschränkung der sensiblen Attribute. Eine *k*-anonyme Gruppe muss zusätzlich mindestens ℓ verschiedene Merkmalsausprägungen innerhalb einer sensiblen Attributklasse aufweisen (Krebs und Hagenweiler 2022). Das verringert zwar den Detaillierungsgrad der Daten, verhindert aber das Auftreten von Gruppierungen, in denen alle Teilnehmer identische sensible Merkmale besitzen (Zheng 2021).

t-Closeness schränkt zudem die Verteilung dieser Merkmale weiter ein. So muss die Variation der Werte innerhalb einer *k*-anonymen Gruppe ähnlich zu der Verteilung im Originaldatensatz sein (Lingala u. a. 2021). Die Variable *t* definiert dabei die erlaubte Abweichung zur ursprünglichen Verteilung.

Diese Ansätze verbessern zwar die *k*-Anonymität, gleichen aber nicht alle Schwächen aus. Da sie die ursprüngliche Überprüfung um weitere Vorausetzungen ergänzt, werden zusätzliche Generalisierungsmaßnahmen benötigt, um das Risiko einer Re-Identifizierung zu vermindern. Das verringert die Datenqualität, Aussagekraft und Verfügbarkeit drastisch. Zudem sind alle drei Verfahren primär für die Anonymisierung von Teildatensätzen geeignet, da sie die Absenz von einzelnen Werten oder deren sinnvolle Zuordnung in eine Gruppe voraussetzen.

Beide Methoden können durch weitere Maßnahmen wie etwa die Aufteilung der Daten in *Cluster* optimiert werden (Lingala u. a. 2021).

C. Differential Privacy

Differential Privacy gehört zur Klasse der statistischen Anonymisierungsverfahren und kommt besonders bei dynamischen Abfragen zum Einsatz (Krebs und Hagenweiler 2022). Ziel des Ansatzes ist es, Antworten nachträglich durch das Hinzufügen von zufällig generiertem Rauschen zu verzerren und einzelne Einträge so unkenntlich zu machen. Die Rauschmenge hängt dabei stark von der Art der Anfrage ab, so benötigen generelle statische Auswertungen im Gegensatz zu stärker personengebundenen Analysen weniger Verzerrungen (Domingo-Ferrer, Sánchez und Blanco-Justicia 2021). *Differential Privacy* zielt dabei nicht darauf ab, den Personenbezug eines einzelnen Eintrags unkenntlich zu machen, sondern dessen Präsenz oder Absenz in einem Datensatz zu verbergen (Domingo-Ferrer, Sánchez und Blanco-Justicia 2021).

Die nachträgliche Verzerrung von Daten findet in der Regel nicht direkt in einer Datenbank oder einer Abfrage statt. In der Praxis übernimmt ein zusätzlicher vertrauenswürdiger Server diesen Anonymisierungsschritt, der als *Middleware*-Komponente arbeitet und die Anfragen entgegennimmt (Krebs und Hagenweiler 2022). Allerdings besteht auch hier die Gefahr, dass Angreifer die dynamischen Anfragen ausnutzen und durch die Verknüpfung der Antwortdatensätze einzelne Personen wieder finden.

Während in den zuvor genannten Verfahren eine Einschränkung auf syntaktischer Ebene erfolgt, versucht *Differential Privacy* eine Re-Identifikation durch zusätzliches Wissen einzuschränken (Krebs und Hagenweiler 2022). Besitzen Dritte ergänzende Kenntnisse über einen (Groß-)Teil der Daten, können sie dies nicht anwenden, wenn nicht ersichtlich ist, ob oder wie viele der bekannten Personen enthalten sind.

Die Bereitstellung von präzisen und trotzdem aussagekräftigen Informationen ist in der Praxis aber nur schwer umsetzbar, da eine detaillierte personenbezogene Datenauswertung dem Grundgedanken der *Differential Privacy* – die Existenz von bestimmten Personen im Datensatz zu verbergen – widerspricht (Domingo-Ferrer, Sánchez und Blanco-Justicia 2021). Die An- und Abwesenheit von bestimmten Personen(-gruppen) hat in derartigen Analysen direkte Auswirkungen auf deren Endresultate.

D. Datensynthese

Die Datensynthese ersetzt persönliche Merkmale durch synthetisch generierte Werte und erzeugt so einen vollständig neuen Datensatz (Raji 2021). Die neuen Charakteristika müssen dabei so gewählt werden, dass jeder künstlich geschaffene Eintrag einer Person im Originaldatensatz entsprechen könnte. Der Synthetisierungsvorgang greift dabei das Konzept der *Differential Privacy* auf (Krebs und Hagenweiler 2022). Eine Teil-Synthetisierung ist ebenfalls möglich (Drechsler und Jentzsch 2017).

Derartige synthetische Einträge können durch das Zusammenspiel zweier neuronaler Netze entstehen. Dabei besitzt ein solches Netz Zugang zu den Originaldaten, während das andere lediglich die syntaktischen Anforderungen sowie die einzelnen Wertebereiche kennt und basierend darauf synthetische Daten erzeugt. Schafft es das erste Netz diese einer realen Person im Datensatz zuzuordnen, muss das zweite den Syntheseprozess überarbeiten (Raji 2021). Synthetische Daten eignen sich für dynamische und statische Abfragen. Bei Ersteren bedarf es allerdings einer stetigen Anpassung des Synthetisierungsmodells (Krebs und Hagenweiler 2022).

Da die erzeugten Datensätze nicht auf reale Personen verweisen, gelten sie nicht mehr als personenbezogene Daten (Raji 2021). Allerdings müssen für ein

gutes Synthetisierungsmodell alle Relationen im Originaldatensatz erkannt und berücksichtigt werden. Durch die Abhängigkeit vom *Differential-Privacy*-Ansatz sinkt die Datenqualität linear mit jedem weiteren Attribut im Datenmodell (Krebs und Hagenweiler 2022). In der Praxis erschweren besonders hochdimensionale Datensätze mit vielen (versteckten) Relationen eine vollständige sowie korrekte Replikation und schränken so die Bereitstellung von Daten ein (Drechsler und Jentzsch 2017).

III. Herausforderung bei der Anonymisierung

Wie bereits an den Schwächen der einzelnen Anonymisierungsverfahren zu erkennen ist, verringern alle Ansätze die Verfügbarkeit von personenbezogenen Daten auf unterschiedliche Arten. Es existiert kein Verfahren, das sowohl maximale Anonymisierung als auch uneingeschränkte Aussagekraft bietet (Drechsler und Jentzsch 2017). Um einen bestmöglichen Anonymisierungsgrad zu erreichen, benötigt es deswegen die Kombination mehrerer Methoden. So kann die *k*-Anonymität die Re-Identifizierung einer einzelnen Person verhindern, während *Differential Privacy* diese Logik ergänzt und dem Wissenszuwachs durch gezielte Datenabfragen entgegenwirkt.

Aus den Schwächen lässt sich zudem schlussfolgern, dass bei der Auswahl von passenden Verfahren drei Kriterien eine wichtige Rolle spielen: die Größe des Datensatzauszuges in Relation zum Originaldatensatz, die Attributvielfalt und die Stärke der nachträglichen Datengeneralisierung sowie -verzerrung. Diese drei Einschränkungen müssen in Einklang miteinander stehen und können nie alle vollständig erfüllt sein (siehe Abbildung 1). Soll der vollständige Datensatz verwendet werden, erfordert das folglich entweder eine starke Reduzierung der Attribute oder deren Genauigkeit.

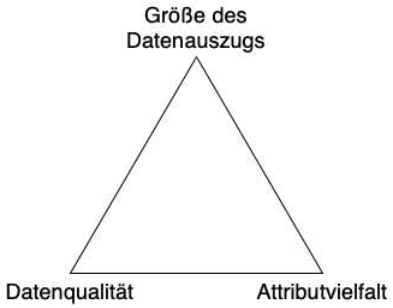

Abbildung 1. Anforderungsdreieck an ein Anonymisierungsverfahren (eigene Darstellung)

Die genannten Anonymisierungsverfahren schränken besonders die Präzision von personenbezogenen Daten stark ein und setzen fast immer Anpassungen der Merkmalsausprägungen voraus, denn werden persönliche Merkmale nur ausgetauscht, handelt es sich lediglich um eine Form der Pseudonymisierung. Diese Gefahr besteht besonders bei synthetischen Daten (Raji 2021).

Des Weiteren basieren alle Ansätze auf unterschiedlichen dynamischen Parametern. Die korrekte Wahl dieser Kenngrößen legt den finalen Grad der Anonymisierung fest. So bedarf es nicht nur der Zusammenführung von mehreren Vorgehensweisen, sondern auch der korrekten Festlegung ihrer Kenngrößen.

Die Anforderungen an den finalen Datensatz und die korrekte Auswahl der Anonymisierungsverfahren sowie deren Parameter haben folglich einen großen Einfluss auf den Detaillierungsgrad und die Menge der verfügbaren Daten.

IV. Fazit und Zusammenfassung

Die vorgestellten Anonymisierungsverfahren lösen zwar das Problem der Verarbeitung personenbezogener Daten, erschweren aber dadurch die Auswertung von großen und gleichzeitig detailreichen Datensätzen. Sie gewährleisten dennoch eine verhältnismäßig sichere Anonymisierung bei weniger umfangreicheren Datenauszügen. Da die Aussagekraft bei mehrdimensionalen Daten abnimmt, reicht es folglich nicht, diese Verfahren auf unveränderte existierende Datenextraktionen anzuwenden.

In stark durch persönliche Daten getriebenen Bereichen wie der medizinischen Forschung oder dem Trainieren von künstlichen Intelligenzen für personenbezogene Prognosen werden deswegen weitere Maßnahmen benötigt. Hier gilt es zunächst festzustellen, welchen Detaillierungsgrad die Auswertungen benötigen, denn das Weglassen irrelevanter Attribute kann schon zu einer höheren Aussagequalität führen. Die einzelnen Analysen verlieren dadurch zwar Zusammenhänge, aber der Verzicht ermöglicht die Verwendung einer größeren Datenmenge. Ein umfangreicher Datensatz mit vielen Wertdopplungen erfüllt beispielsweise Anonymisierungsansätze wie die *k*-Anonymität besser, erfordert weniger Generalisierungsmaßnahmen und erlaubt die Bereitstellung eines größeren Datenauszuges. Ist hingegen die Beständigkeit der Zusammenhänge relevant, verfügen synthetisch generierte Auszüge über mehr Aussagekraft.

Welche Verfahrenskombination die größtmögliche Verfügbarkeit persönlicher Daten sicherstellt, hängt dabei stark vom jeweiligen Anforderungsbereich ab. Dieser legt ebenfalls fest, wie nützlich die Zusammenführung mehrerer Verfahren ist. Hier bedarf es einer Evaluation, ob die Art der personenbezogenen Attribute eine Auswirkung auf die Auswahl der Anonymisierungsverfahren hat.

Literatur

Bieker, F., B. Bremert und M. Hansen (Aug. 2018). „Die Risikobeurteilung nach der DSGVO“. In: *Datenschutz und Datensicherheit* 42.8, S. 492–496.

Domingo-Ferrer, J., D. Sánchez und A. Blanco-Justicia (Juni 2021). „The Limits of Differential Privacy (and Its Misuse in Data Release and Machine Learning)“. In: *Communications of the ACM* 64.7, S. 33–35.

Drechsler, J. und N. Jentzsch (Mai 2017). *Synthetische Daten - Innovationspotential und gesellschaftliche Herausforderungen*. URL: https://www.stiftung-nv.de/sites/default/files/synthetische_daten.pdf (besucht am: 22.11.2022).

Dubli, D. und D. Yadav (Mai 2017). „Secure Techniques for Data Anonymization for Privacy Preservation“. In: *International Journal of Advanced Research in Computer Science* 8.5, S. 1693–1696.

Google LLC (2022). *Google Trends: Anonymisierung und Pseudonymisierung*. URL: https://trends.google.de/trends/explore?date=all&q=%2Fm%2F0277902 (besucht am 22. 11. 2022).

Kesarwani, M. u. a. (2021). „Secure k-Anonymization over Encrypted Databases“. In: *2021 IEEE 14th International Conference on Cloud Computing (CLOUD)* (Chicago, USA, 5.–10. Sep. 2021). IEEE Computer Society, S. 20–30.

Krebs, H.-A. und P. Hagenweiler (2022). *Datenanonymisierung im Kontext von Künstlicher Intelligenz und Big Data*. 1. Aufl. Wiesbaden: Springer Vieweg, S. 91–100.

Li, N., W. Qardaji und D. Su (Juni 2011). *Provably Private Data Anonymization: Or, k-Anonymity Meets Differential Privacy*. Techn. Ber. IN 47907-2086. West Lafayette, Indien: Center for Education, Research - Information Assurance und Security, Purdue University, S. 1–12.

Lingala, T. u. a. (2021). „L-Diversity for Data Analysis: Data Swapping with Customized Clustering“. In: *Journal of Physics: Conference Series* 2089, S. 1–9.

Raji, B. (Jan. 2021). „Rechtliche Bewertung synthetischer Daten für KI-Systeme“. In: *Datenschutz und Datensicherheit* 10.2, S. 303–309.

Zheng, Y. (2021). „Personal Information Protection Based on Big Data“. In: *Journal of Physics: Conference Series* 2037, S. 1–7.