MANZ – QUALITÄT SEIT 1849

02 | 2022 | 93–172 | ecolex

ecolex

FACHZEITSCHRIFT FÜR WIRTSCHAFTSRECHT

Schwerpunkt

Startups und Reform des Gesellschaftsrechts

> Reformbedarf im Aktienrecht
> Für und Wider eine neue Rechtsform
> Startups und Mitarbeiterbeteiligung

Digitale Leistungen: Ausschluss der Aktualisierungspflicht

Datenanonymisierung als Schlüssel zur Innovation?

EuGH-Update zur Umsatzsteuer

Zur Verfassung von Erneuerbare-Energie-Gemeinschaften

Keine Geschäftsführerhaftung bei Cyber-Attacke

UrhG-Nov 2021: Zur Umsetzung der Online-SatKab-RL

ECOLEX.MANZ.AT

ISSN 1022-9418 Österreichische Post AG MZ 02Z032706 M Verlag Manz, Gutheil Schoder Gasse 17, 1230 Wien

MANZ

Datenanonymisierung – Der Schlüssel zur Innovation

Herausforderungen und Lösungswege in der Praxis

BEITRAG. Um Chancen für Innovationen zu eröffnen und die oft herrschenden komplexen datenschutzrechtlichen Hürden zu überwinden, setzen Unternehmer vermehrt auf die Datenanonymisierung. Aber auch dabei stellen sich rechtliche und technische Fragen rund um die Effektivität der Anonymisierung.
ecolex 2022/114

Georg Heiler, MSc., ist wissenschaftlicher Mitarbeiter der TU Wien und des Complexity Science Hub Vienna und Data Scientist bei der T-Mobile Austria GmbH.
Mag. **Alexandra Ciarnau** ist Rechtsanwältin im IP/IT- und Datenschutzteam bei DORDA Rechtsanwälte GmbH und Co-Leiterin der Digital Industries Group.

A. Einleitung

Neue Technologien, wie etwa künstliche Intelligenz, autonomes Fahren oder Smart Homes, setzen allesamt eine Fülle an Daten voraus, um ihr Potential vollends entfalten zu können. Dabei ist ein Personenbezug häufig ein nicht notwendiger Nebeneffekt. Die damit einhergehenden komplexen datenschutzrechtlichen Fragen wirken oft innovationshemmend und verlangsamen Projekte. In der Praxis ist daher die Datenanonymisierung eine attraktive Lösung, um beide Welten – den Datenschutz und den digitalen Fortschritt – einfach miteinander zu vereinbaren. Die DSGVO und das DSG sind schließlich bei Verarbeitung anonymer Daten nicht anwendbar. Eine vollständige Anonymisierung, die jeglichen Personenbezug entfernt, ist aber schwer zu erzielen. Das gilt insb dann, wenn viele Daten kombiniert werden und möglichst viel Informationsgehalt erhalten bleiben soll.[1)] Schließlich können nur aussagekräftige Informationen gewinnbringend genutzt werden. Dieses Dilemma beschäftigt auch die Statistik:

Klassische statistische Modelle reduzieren oftmals den Informationsgehalt der Daten sehr stark oder anonymisieren nicht ausreichend. Daher sind diese Techniken nicht lukrativ. Datensynthetisierung bzw homomorphe Kryptographie sind hingegen in komplexen Ausnahmefällen zielführender. Beide Methoden sind aber wesentlich komplexer und weisen einen hohen Energieverbrauch auf. Ein Verfahren, das auf der probabilistischen Anonymisierung beruht, ist jedoch für eine Vielzahl von Anwendungsfällen geeignet, einfach verständlich und energieeffizient umsetzbar. In der Praxis haben sich daher probabilistische Anonymisierungsmethoden durchgesetzt. Dadurch soll die Balance zwischen dem Informationsgehalt der Daten und Anonymität erreicht werden. Im Folgenden geben wir einen Überblick über die rechtlichen Voraussetzungen und den Stand der Technik der probabilistischen Anonymisierung.

B. Anonymisierung – rechtliche Grundlagen

1. Definition

Wie bereits erwähnt, ist die DSGVO nicht auf anonyme Daten anwendbar.[2)] Damit greifen in diesem Fall auch nicht die strengen datenschutzrechtlichen Pflichten. Fraglich ist allerdings, was unter „Anonymisierung" zu verstehen ist. Die DSGVO definiert diesen Begriff zwar nicht, geht aber in ErwGr 26 DSGVO darauf ein. Demnach sind anonymisierte Daten Informationen ohne Personenbezug iSd Art 4 Z 1 DSGVO. Es ist daher stets im Einzelfall zu prüfen, ob ein Rückschluss auf eine natürliche Person über das Datum bzw durch Verknüpfung mit weiteren Informationen möglich ist.[3)] Als Beispiele einer direkten oder indirekten Identifizierung nennt Art 4 Z 1 DSGVO die *„Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung, oder zu einem oder mehreren besonderen Merkmalen, die Ausdruck der physischen, physiologischen, genetischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind"*.[4)] Der Personenbezug kann über unterschiedliche Informationen erfolgen. Auch wenn eine Kennung für sich alleine die Ermittlung einer Person nicht ermöglicht, kann die Identifizierbarkeit auch über eine Kombination mehrerer Elemente möglich sein.[5)]

[1)] *Winter/Battis/Halvani,* Herausforderungen für die Anonymisierung von Daten, ZD 2019, 489.

[2)] E contrario Art 4 Z 1 DSGVO.

[3)] *Klar/Kühling* in *Kühling/Buchner* (Hrsg), DS-GVO BDSG² (2018) Art 4 Rz 19.

[4)] Dabei sind alle Mittel zu berücksichtigen, die vom Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um Personen zu identifizieren. Vgl hierzu ErwGr 26 DSGVO; *Klar/Kühling* in *Kühling/Buchner,* DS-GVO BDSG² Art 4 Rz 21.

[5)] *Ziehbarth* in *Sydow,* Europäische Datenschutzgrundverordnung² (2018) Art 4 Rz 17 f; vgl auch De-Anonymisierung von Daten mit inhärenter Struktur wie bspw bei sozialen Netzwerken oder Geodaten, *Shouling Ji/Weiqing Li/Mudhakar Srivatsa/Jing Selena He/Raheem Beya,* Structure Based Data De-Anonymization of Social Networks and Mobility Traces, Information Security (2014) 237 ff.

2. Anonymisierungsgrad

Aus praktischer Sicht spannend ist daher der Grad der notwendigen Anonymisierung. Dabei wird zwischen der absoluten und faktischen Anonymisierung unterschieden:

Im ersten Fall wird jedes potenzielle Identifikationskriterium neutralisiert, sodass niemand mehr in der Lage ist, den Personenbezug herzustellen. Das ist die verlässlichste Variante. Sie wird vor allem bei veröffentlichten amtlichen Statistiken angewandt (zB Wahlergebnisse). Sie erlauben keinen Rückschluss auf die ihnen zugrunde liegenden personenbezogenen Einzelangaben mehr.

Das Datenschutzrecht verlangt aber keine absolute Anonymisierung. Es begnügt sich mit der faktischen Anonymisierung (auch „relative Anonymisierung" genannt). Dabei werden so viele Merkmale entfernt, dass die Identifizierung nach dem Stand der Technik nicht mehr mit vernünftigerweise zu erwartendem Aufwand und absolut verlässlich erreicht werden kann. Die faktische Anonymisierung basiert daher auf einer Risikoprognose.[6] Je nach Konstellation kann sie aber auch ausreichen, um die absolute Anonymität zu erreichen.[7] In diese Kerbe schlägt auch die in der Praxis beliebte Form der probabilistischen Anonymisierung. Sie geht von einer hohen Wahrscheinlichkeit der fehlenden Rückverfolgbarkeit aus.

3. Abgrenzung zur Pseudonymisierung

Die faktische Anonymisierung scheint auf den ersten Blick Gemeinsamkeiten mit der Pseudonymisierung zu haben. Tatsächlich handelt es sich dabei aber um zwei verschiedene Paar Schuhe: Nach Art 4 Z 5 DSGVO ist unter Pseudonymisierung die Verarbeitung personenbezogener Daten zu verstehen, die durch Hinzuziehung zusätzlicher Informationen doch noch einer spezifischen betroffenen Person zugeordnet werden können. Die Identifikation wird daher nicht vollständig beseitigt, sondern kann vom Verantwortlichen jederzeit wiederhergestellt werden.[8] Diese technische Methode führt noch nicht zur Unanwendbarkeit der DSGVO. Sie ist vielmehr als technische und organisatorische Maßnahme zum Schutz der Geheimhaltungsinteressen der Betroffenen zu betrachten.

C. Praktische Lösungsansätze der Anonymisierung

Für Verantwortliche ist interessant, wie sie möglichst viel Informationsgehalt der Daten bei maximal möglicher Anonymisierung erreichen können. In der Privatwirtschaft wird daher die probabilistische Anonymisierung bevorzugt. Hierfür gibt es zwei etablierte Konzepte, die miteinander kombiniert werden können: (a) K-Anonymität und (b) Differential Privacy.[9] Dazu im Detail:

1. Die K-Anonymität

Mithilfe der K-Anonymität können Aussagen über anonymisierte Datensätze getroffen werden. Aus mathematischer Sicht muss hierfür bei der K-Anonymität eine Mindestanzahl von „K" (zB 10) nicht unterscheidbaren Individuen in eine Gruppe fallen, die aus den Datenfeldern definiert wird. Das bedeutet, dass entscheidende Informationen verloren gehen müssen, um die Garantien der K-Anonymität einzuhalten. Das lässt sich anhand einer Tabelle gut veranschaulichen, die Identifikatoren (zB Name), Quasi-Identifikatoren (zB Alter, Geschlecht) und sonstige Attribute (zB Krankheit, Hobbys, Berufe etc) enthält. Die Identifikatoren werden entfernt und die Personen anhand der Quasi-Identifikatoren gruppiert (zB 10 Pax/50 Pax im Alter zwischen 40 bis 50 Jahren/männlich leiden unter Haarausfall).

2. Differential Privacy

Durch vorausgehendes Wissen über den spezifischen Datensatz kann aber potenziell eine De-Anonymisierung von Individuen erfolgen. Deswegen setzt Differential Privacy eine wiederholte Auswertung der gleichen Daten voraus, die nicht zu dem exakt gleichen Ergebnis führen dürfen.[10] Dieses Verfahren mischt daher den echten Daten ein zufälliges Rauschen bei.

3. Wesensmerkmale und Potentiale beider Methoden

Beide Verfahren zerstören für sich allein die Daten unnötig stark, um die entsprechenden Garantien der Anonymisierung zu gewährleisten. Stattdessen können aber beide Verfahren miteinander kombiniert werden. Dadurch weist die geballte Anonymisierung folgende Eigenschaften auf:

- Der Grad der Anonymisierung ist einstellbar (K kann je nach Bedarf gewählt werden);
- es wird genauso viel Rauschen wie benötigt hinzugefügt;
- die anonymisierten Daten sind mathematisch garantiert.

4. Praxisbeispiel

Folgendes Beispiel verdeutlicht die kombinierte Anonymisierungsmethode:

Für eine Identität bzw einen Primärschlüssel soll ein Datensatz anonymisiert werden.[11] Hierfür gehen wir zunächst von einer K-Anonymität von zehn aus. Das bedeutet, dass mindestens zehn unterschiedliche Identitäten in die gleiche Gruppe aus Datenpunkten fallen müssen. Folglich kann mit einer Wahrscheinlichkeit von 1/10 = 10% zufällig ein Datenpunkt de-anonymisiert werden.

Zusätzlich werden die Daten mit einem Rauschen versehen.[12] Das erfolgt aber nur innerhalb der Nachbarschaft, um eine möglichst gute Qualität beibehalten zu können. Wenn wir *Gauss*-basiertes Rauschen einsetzen, kann sogar ein höherer Anonymisierungsgrad erzielt werden. Mit einem multidimensionalen Index lässt sich die Nachbarschaft einfach berechnen.

---

[6] *Roßnagel*, Datenlöschung und Anonymisierung, ZD 2021, 188 (189).

[7] *Ziehbarth* in *Sydow*, Europäische Datenschutzgrundverordnung (2018) Art 4 Rz 30. Was heute möglicherweise für eine faktische Anonymisierung ausreicht, kann aber morgen bereits veraltet sein. Mit dem Zuwachs an neuen Technologien und dem Fortschritt wie bspw Quantencomputing könnte eine De-Anonymisierung zB effizienter und verlässlicher erreicht werden.

[8] *Weth/Herberger/Wächter/Sorge*, Daten- und Persönlichkeitsschutz im Arbeitsverhältnis[2] (2019).

[9] *Li/Qardaji/Su*, Provably Private Data Anonymization: Or, k-Anonymity Meets Differential Privacy, https://dblp.org/rec/journals/corr/abs-1101-2604.html (abgerufen am 17. 1. 2022); *Emam/Dankar*, Protecting Privacy Using k-Anonymity, https://academic.oup.com/jamia/article/15/5/627/732733?login=true (abgerufen am 17. 1. 2022).

[10] *Smith*, Pinning Down „Privacy" in Statistical Databases, https://cs-people.bu.edu/ads22/talks/2012-08-21-crypto-tutorial.pdf (abgerufen am 17. 1. 2022).

[11] Der Einfachheit halber generieren wir einen zweidimensionalen, zufällig erzeugten Datensatz.

[12] Künstliches Rauschen nach *Gauss* wird zu den Daten hinzugefügt, um die ursprünglichen Beobachtungen nicht deterministisch zu verändern.

Diese Schritte könnten in der Praxis wie folgt leicht umgesetzt werden:[13)]

- Zunächst müssen Beispieldaten erzeugt werden (zB ein zweidimensionaler Datensatz pro Identität, vgl „lat1" und „long1"):

	id_str	lat1	long1
0	0	0.548814	0.715189
1	1	0.602763	0.544883
2	2	0.423655	0.645894

Abbildung 1

- Anschließend wird die K-Nachbarschaft berechnet. Hierfür kann ein multidimensionaler Index genutzt werden, um die Berechnung effizient durchführen zu können. Zu den ursprünglichen zwei Attributen werden daher alle K=10 nächsten Nachbarn (temporär) mitabgespeichert und die maximale Distanz zwischen der aktuellen Identität und den 10 Kandidaten berechnet:

id_str	lat1	long1	complex_type	dist_same_order	max_dista
0	0.548814	0.715189	[[0.46147936225293185, 0.7805291762864555], [0.6120957227224214, 0.6169339968747569], [0.4236547993389047, 0.6458941130666561], [0.45615033221654855, 0.5684339488686485], [0.6027633760716439, 0.5448831829968969], [0.4375872112626925, 0.8917730007820798], [0.5680445610939323, 0.925596638292661], [0.7781567509498505, 0.8700121482468192], [0.26455561210462697, 0.7742336894342167], [0.521848321750...	[0.10907127514430776, 0.1168706843085668, 0.14306129268590195, 0.17356156244458887, 0.1786470956951876, 0.20869371845180915, 0.21128429576441266, 0.2767099903186786, 0.2903253022891915, 0.3017347428745624]	0.301735

Abbildung 2

- Da stets von einer Worst-Case-Annahme ausgegangen werden muss, wird die Maximaldistanz als Standardabweichung genutzt, um das minimal zu verwendende Rauschen zu definieren. Das Ergebnis vergleicht die originalen Daten direkt mit den frisch erzeugten (anonymen) Attributen:

	id_str	lat1	lat1_sythesized	max_distance	long1	long1_sythesized
0	0	0.548814	0.715189	0.301735	1.038935	0.383094
1	1	0.602763	0.544883	0.382293	0.368893	0.982503
2	2	0.423655	0.645894	0.314773	0.257401	0.929690
3	3	0.437587	0.891773	0.384208	0.025344	1.084835
4	4	0.963663	0.383442	0.600408	1.483261	0.924323

Abbildung 3

- Zu jedem originalen Datenpunkt ist jetzt entsprechend den Anonymitätsanforderungen ein passendes Tupel anonymer Attribute generiert worden.
- Validierung der Korrektheit: In unserem Beispiel gehen wir davon aus, dass eine De-Anonymisierung im Worst Case auftreten könnte, wenn der erste (= nächste Nachbar) dem echten originalen Datenpunkt entspricht. Es ist daher wesentlich, dass ein einzelner Durchlauf zusätzlich den Anforderungen an die Anonymität genügt, eine wiederholte Anonymisierung (mit bspw täglich frischen Daten) auch im Zeitverlauf anonym ist und trotz gleicher Daten nicht die exakt gleichen Ergebnisse liefert.
- Für 80% der Daten muss bspw nur eine geringe Menge an zusätzlichem Rauschen hinzugefügt werden, um die Daten entsprechend zu anonymisieren. Die Distanz zu den echten Punkten ist nur so groß wie eben notwendig, um die gewünschte Garantie bzw Qualität der Anonymisierung zu erreichen. Dies zeigt sich im Plot der Quantile (x-Achse) und deren jeweils notwendiger Menge an Rauschen (y-Achse):

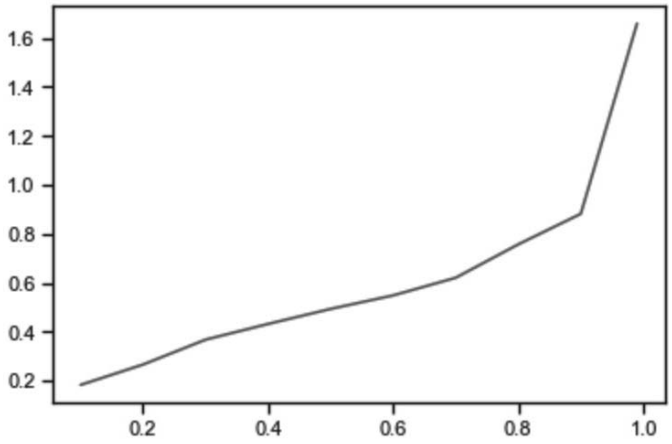

Abbildung 4

Schlussstrich

Die kombinierte Methode der K-Anonymität und Differential Privacy eignen sich aufgrund ihrer Effizienz und Nachvollziehbarkeit für viele praktische Anwendungsfälle. Zudem kann das Verfahren auch je nach konkreter Ausgestaltung und Hinzufügen des Rauschens den hohen rechtlichen Anforderungen der Anonymität iSd DSGVO genügen. Datenbasierte Produkte, die diese Methoden nutzen, können daher einfacher datenschutzkonform eingesetzt werden.

[13)] Ein komplettes Code-Beispiel mit weiteren Erklärungen findet sich auch unter https://georgheiler.com/2021/03/08/can-you-tell-the-nuts-berries-apart-in-each-group (abgerufen am 17. 1. 2022).