Institut für Parallele und Verteilte Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Masterarbeit

Anonymisierung von Daten: Von der Literatur zum Automobilbereich

Jan Herkommer

Studiengang:	Informatik
Prüfer/in:	Prof. Dr. Bernhard Mitschang
Betreuer/in:	Dr. rer. nat. Pascal Hirmer, Andrea Fieschi, M.Sc.
Beginn am:	28. September 2022
Beendet am:	28. März 2023

Kurzfassung

Die Datenanonymisierung im Automobilbereich gewinnt immer mehr an Bedeutung. Jedoch gibt es kaum Literatur und Ansätze, die sich mit der Anonymisierung von Automobildaten beschäftigen. In dieser Arbeit werden deshalb mit Hilfe einer strukturierten Literaturrecherche die aktuell verbreitetsten Verfahren und Anwendungsbereiche erörtert und die wichtigsten Erkenntnisse der Recherche zusammengefasst. So werden bei den analysierten Paper der Anwendungsbereich, die Methodik sowie der zu anonymisierende Datentyp ermittelt. Des Weiteren werden die Metriken zum Vergleich von unterschiedlichen Ansätzen betrachtet. Mit Hilfe dieser Erkenntnisse wird im Anschluss auf die Anonymisierung von Fahrzeugdaten anhand verschiedener Anwendungsfälle eingegangen und Herausforderungen und Lösungsansätze skizziert. Zuletzt wird beispielhaft ein Ansatz zur Anonymisierung von Routen implementiert, um mit Hilfe eines GPS-Sensors aufgezeichnete Fahrzeugrouten zu anonymisieren. Dabei werden zusätzliche Probleme wie der Umgang mit Messungenauigkeiten und Messfehlern sowie die tatsächlichen Auswirkungen von reduzierter Datennutzbarkeit verdeutlicht.

Inhaltsverzeichnis

1 Einleitung **13**

2 Grundlagen der Datenanonymisierung **15**
2.1 Definition von Anonymisierung 15
2.2 Ansätze von Anonymisierung 17

3 Systematische Literaturrecherche **23**
3.1 Vorgehen 23
3.2 Ergebnisse 25

4 Transfer zur Automobilwelt **35**
4.1 Anwendungsfälle in der Automobilwelt 35
4.2 Weitere Herausforderungen 38

5 Implementierung ausgewählter Ansätze **41**
5.1 Datengenerierung 41
5.2 Auswahl des Papers 43
5.3 Implementierung 44

6 Ergebnisse **59**

7 Ausblick **61**

Literaturverzeichnis **63**

Abbildungsverzeichnis

2.1 Abhängigkeit von Informationsgehalt und Stärke der Anonymität [CNV+21] 17
2.2 Beispieldaten ohne Anonymisierung 18
2.3 k-anonyme Daten mit k=2 (2-anonymity) 19
2.4 l-diverse Daten mit l=2 (2-diversity). Außerdem wurde t-closeness (mit t=0,33) auf diese Daten angewendet 20
2.5 Noch stärker verallgemeinerte Daten mit k=4, l=3 und t=0,19 21

3.1 Anzahl der gefundenen Dokumente mit verschiedenen Suchstrings auf scopus 24
3.2 Einzelne Schritte bei der strukturierten Literaturrecherche nach dem PRISMA-Ansatz. 25
3.3 Verteilung der verschiedenen Methodiken in den analysierten Arbeiten. . 27
3.4 Anwendungsgebiete der analysierten Arbeiten mit der jeweiligen Verteilung der Methodiken in diesem Gebiet. 29
3.5 Datentypen der analysierten Arbeiten mit der jeweiligen Verteilung der Methodiken in diesem Gebiet. 30

5.1 Visualisierung der mit Hilfe des On-Board Device (OBD) GPS-Trackers aufgezeichneten Routen in einer Woche. 43
5.2 Ausreißer in den aufgezeichneten GPS-Daten 46
5.3 width=0.6 50
5.4 Links: Nur Anonymisierung des sensitiven Punkts. Rechts: Anonymisierung aller Punkte, die im Minimum Bounding Rectangle (MBR) liegen . . 51
5.5 Links: Original Route. Mitte: Anonymisierte Route mit Zentrum des MBR. Rechts: Anonymisierte Route ohne Zentrum des MBR 53
5.6 Links: Original Route. Rechts: Anonymisierte Route mit MBR 54
5.7 Links: Map-Matching um gemessene Positionen an Verkehrswege zu binden. Abbildung aus [BK96]. Rechts: Gemessener Punkt mit dem OBD 56

Verzeichnis der Algorithmen

5.1 Datenaufbereitung 45
5.2 Anonymisierung 47
5.3 Berechnung Haltingpoints 48
5.4 Berchenung Punkte für MBRs 49
5.5 Berechnung MBRs 51
5.6 Ersetzen aller Punkte in MBRs 52

Abkürzungsverzeichnis

GAN Generative Adversarial Networks. 28

GDPR General Data Protection Regulation. 15

MBR Minimum Bounding Rectangle. 7

OBD On-Board Device. 7

OSN Online Social Networks. 29

PII persönlich identifizierbare Informationen. 15

POI Point-Of-Interest. 30

1 Einleitung

Bis zu 600 Gigabyte Daten werden in den neuesten Autos laut einem führenden Unternehmen zur Verarbeitung von Fahrzeugdaten (Harmnan International) pro Tag generiert [Lit]. Zwar wird nur ein Bruchteil davon weiterverarbeitet, jedoch zeigt die Höhe der Zahl die Dimensionen, in denen sich Autohersteller inzwischen bewegen. Die alleinige Aufzeichnung der Daten bringt jedoch zunächst keinen Mehrwert. Für die Automobilhersteller ist es von großer Bedeutung, diese Daten - oder zumindest einen Teil davon - auszuwerten, um somit neue Erkenntnisse aus diesen Daten zu gewinnen. So können die Daten zur Verbesserung verschiedener Bauteile (Batterien, Bremsen, Schaltung, ...) genutzt werden. Es können jedoch auch Echtzeitdaten genutzt werden, um Informationen über die aktuelle Verkehrslage zu gewinnen oder sogar Informationen über Straßen gesammelt werden. Nicht zuletzt für die Entwicklung im Bereich Autonomes Fahren dürfte die Nutzung und Erforschung verschiedenster Daten an Bord der Autos enorm wichtig sein.

Die Daten sind jedoch zwangsläufig zunächst mit dem Automobil, das die Daten aufzeichnet, und damit mit dessen Inhaber verknüpft. Was bei Daten zur Motordrehzahl zunächst unbedenklich erscheint, wird spätestens bei geografischen Daten zum Problem. Zum einen muss es im Interesse der Automobilhersteller sein, die Daten der Kunden nicht preiszugeben, zum anderen sind es Vorgaben der Europäischen Union zum Datenschutz, die die Hersteller zwingen, die aufgezeichneten Daten ausreichend zu schützen.

Um die Daten trotzdem nutzen zu können, müssen diese anonymisiert werden, sodass sie nicht mehr einzelnen Personen oder Autos zuordenbar sind. Während Datenanonymisierung in anderen Bereichen, zum Beispiel in der Medizin bei klinischen Studien, bereits seit langer Zeit Alltag ist, ist die Anonymisierung im Automobilbereich bis jetzt kaum ein Thema gewesen. Erst die zunehmende Verfügbarkeit der Daten macht das Thema Anonymisierung auch im Automobilbereich wichtig. Aktuell gibt es jedoch kaum Studien und Ansätze zur Anonymisierung von Automobildaten. Um einen Überblick über aktuell gängige Anonymisierungsverfahren in anderen Bereichen sowie neueste Entwicklungen und Trends bei der Anonymisierung von Daten zu bekommen, soll in dieser Arbeit eine umfassende Literaturrecherche zum Thema Anonymisierung erfolgen. Neben der Darstellung der aktuell vorherrschenden Verfahren und Methodiken sollen zudem Ideen geschildert werden, für welche Bereiche die verschiedenen Verfahren für Automobildaten relevant sein können. So werden anhand verschiedener Anwendungsfälle die Erkenntnisse der Literaturrecherche auf den Automobilbereich übertragen. Zudem wird auf weitere Herausforderungen im Rahmen der Datenanonymisierung im Automobilbereich eingegangen.

Um eine Aussage über die grundsätzlichen Möglichkeiten zur Nutzung der in der Ausarbeitung dargestellten Ansätze treffen zu können, sollen simulierte Daten aus dem Automobilbereich genutzt werden und mit Hilfe ausgewählter Ansätze anonymisiert werden. Hierbei geht es weniger um eine umfassende Implementierung, um die Daten vollständig zu anonymisieren als vielmehr um einen ersten Versuch, Ansätze anderer Anwendungsbereiche auch für Automobildaten zu nutzen. Dazu wurde ein Ansatz von Rajesh et al. aus dem Jahr 2019 [RAD19] genutzt, um mit Hilfe eines GPS-Sensors aufgezeichnete Daten aus Fahrzeugen zu anonymisieren.

Die Arbeit ist wie folgt aufgebaut. Zunächst wird in Kapitel 2 der Begriff Anonymisierung genauer definiert und auf die grundsätzlichen Möglichkeiten zur Anonymisierung von Daten eingegangen. In Kapitel 3 werden die Vorgehensweise und die Ergebnisse der Literaturrecherche dargestellt. Anschließend werden in Kapitel 4 die Erkenntnisse der Literaturrecherche anhand verschiedener Anwendungsfälle im Automobilbereich betrachtet und weitere Herausforderungen der Datenanonymisierung in diesem Bereich erörtert. Zuletzt wird in Kapitel 5 eine beispielhafte Implementierung erläutert, ehe am Ende in Kapitel 6 ein Fazit gezogen wird.

2 Grundlagen der Datenanonymisierung

Um sich den verschiedenen Vorgehensweisen und Ansätzen zur Anonymisierung von Daten zu widmen, soll im Folgenden zunächst auf die Grundlagen der Datenanonymisierung eingegangen werden. Zuerst wird der Begriff Anonymisierung genauer definiert und abgegrenzt. Anschließend sollen zwei wichtige Grundkonzepte der Datenanonymisierung kurz erläutert werden, auf die im Verlaufe der Arbeit hingewiesen wird.

2.1 Definition von Anonymisierung

Der Begriff Anonymisierung ist weithin bekannt und im Allgemeinen wird darunter vereinfacht gesagt die Verfremdung von Daten verstanden. Anonymisierung näher zu definieren fällt jedoch den meisten Personen bereits deutlich schwerer.

Wirft man einen Blick in den Duden, so wird dort der Begriff „anonymisieren“ (Anonymisierung als Substantiv gibt es nicht) folgendermaßen beschrieben: „(aus einer Statistik, aus Fragebogen, Unterlagen u. Ä.) den Namen einer Person, persönliche Daten löschen“ [Dud]. Diese allgemeine Definition von Anonymisierung beschreibt das Vorgehen relativ gut, lässt jedoch noch viel Spielraum für Interpretationen was zum Beispiel mit „persönliche Daten“ gemeint ist.

Um den Begriff Anonymisierung (engl. Anonymization) im Zusammenhang mit Daten zu definieren, hilft ein Blick in den ISO-Standard (ISO/IEC 29100:2011) für Information Technology zu Privacy und Datenschutz. Laut diesem beschreibt Anonymisierung den „Prozess in dem persönlich identifizierbare Informationen (PII) unwiderruflich verändert werden, sodass die zugehörige Person nicht mehr direkt oder indirekt identifiziert werden kann. Weder vom Inhaber der PII alleine, noch mit Unterstützung anderer Quellen.“ [11]. Die Definition macht deutlich, dass es zunächst darum geht, sämtliche Informationen zu verändern, die auf die zugehörige Person schließen lassen. Hierbei spielt das Wort „unwiderruflich“ eine große Rolle, da damit eine Abgrenzung zu anderen Verfahren wie Pseudonymisierung oder Verschlüsselungsverfahren (siehe im Folgenden) gezogen wird.

Eine deutliche kürzere Definition gibt die General Data Protection Regulation (GDPR). Diese EU-Richtlinie zu Datensicherheit und Datenschutz gibt unter anderem Aufschluss über die Vorgaben bezüglich der Verwendung personalisierter Daten und gilt in der gesamten

Europäischen Union. Im Recital 26 der GDPR wird beschrieben, für welche Daten die GDPR-Vorgaben eingehalten werden müssen. Hierbei werden anonymisierte Daten explizit ausgenommen. Anonymisierte Informationen sind laut Recital 26 „... Informationen, die sich nicht auf eine bestimmte oder bestimmbare natürliche Person beziehen, oder personenbezogene Daten, die so anonymisiert wurden, dass die betroffene Person nicht oder nicht mehr identifizierbar ist.“ [Vol22]. Zudem findet eine klare Abgrenzung zu pseudonymisierten Daten statt. Pseudonymisierung ist laut ISO-Standard ein „... Verfahren, das auf PII angewandt wird und bei dem identifizierende Informationen durch einen Alias ersetzt werden“ [11]. Bei pseudonymisierten Daten ist es sowohl für den Inhaber der Daten, als auch für externe Personen mit Hilfe zusätzlicher Daten möglich, die ursprüngliche Identität wieder herzustellen.

Eine Handreichung zur GDPR präzisiert den Begriff Anonymisierung und die Abgrenzung zu anderen Verfahren weiter. So wird in [Sup] neben Pseudonymisierung auch Verschlüsselung von Anonymisierung abgrenzt, da diese normalerweise durch die passenden geheimen Schlüssel wieder rückgängig gemacht werden kann und damit die Veränderung der Daten nicht irreversibel ist. Zudem wird eine erste Idee über die Messbarkeit von Anonymisierung gegeben. Mit der Re-Identifizierungswahrscheinlichkeit wird die Wahrscheinlichkeit beschrieben, dass anonymisierte Daten durch verschiedene Verfahren wieder zu personenbezogenen Daten umgewandelt werden können. Jedoch wird nicht näher auf die Re-Identifizierungswahrscheinlichkeit (engl. „Re-Identification Likelihood“) eingegangen. Auch findet sich in der Literatur dieser Begriff nirgends.

In der Handreichung wird ein weiterer sehr wichtiger Punkt bezüglich Anonymisierung hervorgehoben. Der Anonymisierungsprozess ist immer eine Abwägung zwischen der Minimierung der Re-Identifizierungswahrscheinlichkeit und der Möglichkeit zur weiteren Nutzung der anonymisierten Daten (utility). Es ist möglich, Daten vollständig zu anonymisieren. Diese Daten können jedoch für eine sinnvolle Analyse nicht mehr genutzt werden. Auf der anderen Seite können Daten, die gar nicht anonymisiert werden, sehr gut für Analysen genutzt werden. Allerdings sind alle Daten dann persönlich identifizierbar.

Abbildung 2.1 zeigt die Abhängigkeit von Informationsgehalt der Daten und Anonymität. Pseudonymisierte Daten haben ein geringes Anonymitätslevel, enthalten durch die geringe Veränderung der Daten aber mehr Informationen. Je stärker Daten anonymisiert werden, desto geringer wird der Informationsgehalt der Daten. Ziel bei der Anonymisierung ist es grundsätzlich, eine Balance zwischen Informationsgehalt und Anonymitätslevel zu finden.

Ein letzter Punkt, auf den in [Sup] hingewiesen wird, betrifft die Endgültigkeit von Anonymisierung. Ein anonymisierte Datensatz ist nicht zwangsläufig für immer anonym. Durch fortschreitende Technik, mehr Rechenleistung und ggfs. auch neuen Quellen können Daten später re-identifiziert werden, obwohl sie zum Zeitpunkt der Anonymisierung noch als anonym gegolten haben. Dies betrifft vor allem Daten, deren Re-Identifizierungswahrscheinlichkeit von Beginn an etwas höher war, um einen höheren Informationsgehalt zu erlangen.

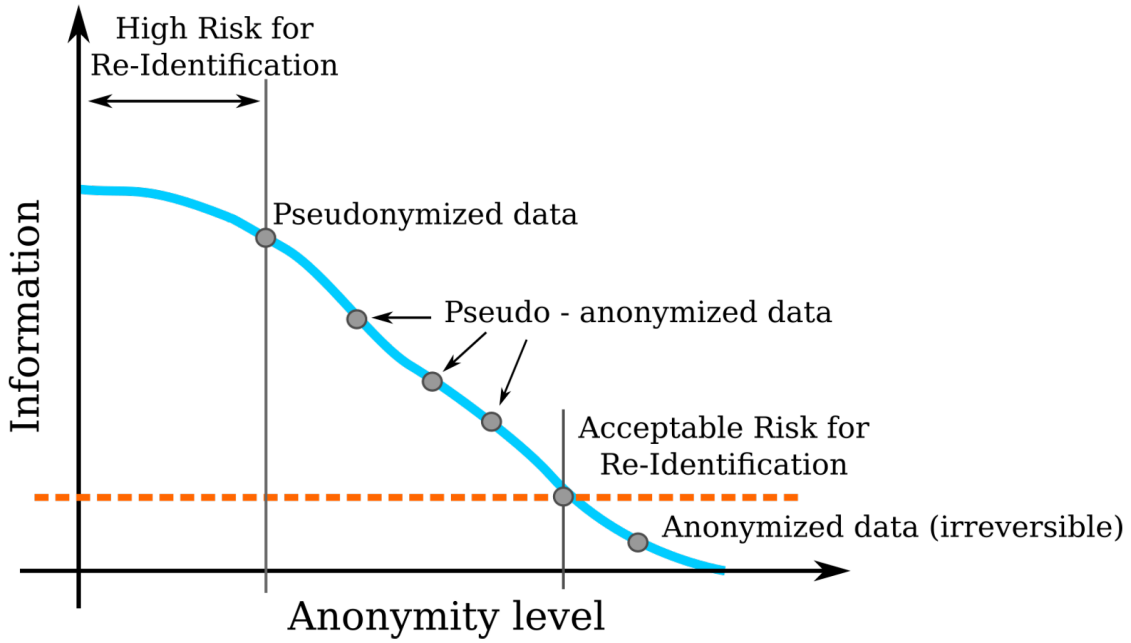

Abbildung 2.1: Abhängigkeit von Informationsgehalt und Stärke der Anonymität [CNV+21]

2.2 Ansätze von Anonymisierung

Grundsätzlich gibt es verschiedene Vorgehensweisen, wie Daten anonymisiert werden können. Bei den pertubativen Verfahren (Randomisierung) werden die Daten mit Hilfe mathematischer Verfahren verändert. Es wird Rauschen hinzugefügt und Werte verändert, sodass die Daten nicht mehr eindeutig auf Personen bezogen werden können. Zwar wird die Wahrscheinlich einer Re-Identifizierung damit verringert, es reduziert sich allerdings auch die Nutzbarkeit der Daten, da die ursprünglichen Werte verändert wurden.

Bei der Generalisierung werden die Daten nicht willkürlich verändert, sondern verallgemeinert. Es wird versucht, die einzelnen Datensätze nicht mehr eindeutig identifizierbar zu machen. Zum Beispiel können bei der Postleitzahl nur die ersten drei Ziffern verwendet werden. Die Informationen entsprechen nach wie vor der Wahrheit, sind aber nicht mehr so genau. Des Weiteren werden einzelne Werte, wenn sie nicht oder nicht ausreichend generalisierbar sind, unterdrückt, also weggelassen. Auch bei diesem Verfahren reduziert sich die Nutzbarkeit der Daten. Außerdem ist hier der anti-proportionale Zusammenhang zwischen Re-Identifizierungswahrscheinlichkeit und Nutzbarkeit der Daten besonders gut zu sehen. Je allgemeiner die Daten generalisiert werden, desto höher ist der Schutz vor Re-Identifizierung und desto niedriger ist die Nutzbarkeit.

Die meisten Ansätze zur Anonymisierung von Daten lassen sich auf diese beiden Verfahren zurückführen. Sweeney hat im Jahr 2002 mit dem Konzept der k-anonymity [Swe02] die Grundlage für aggregationsbasierte Verfahren gelegt. Dies soll im Folgenden kurz

vorgestellt werden. Als wichtiger Vertreter von pertubativen Verfahren gilt das Modell der Differential Privacy von Dwork aus dem Jahr 2006 [Dwo08]. Dieser Ansatz wird anschließend ebenfalls kurz vorgestellt.

2.2.1 k-anonymity

Die grundsätzliche Idee bei k-anonymity [Swe02] besteht darin, Daten so zu verändern, dass es immer mindestens k gleiche Einträge gibt, die eine eindeutige Identifizierung eines Datensatzes verhindern. Dazu werden die Quasi-Identifier, also Attribute eines Datensatzes, die indirekt im Zusammenhang mit anderen Attributen auf die zugehörige Person bezogen werden können, generalisiert. Je höher k dabei gewählt wird, desto stärker ist die Anonymisierung.

Identifier	Quasi-Identifier		Sensitives Attribut
ID	**Alter**	**PLZ**	**Krankheit**
1	25	70368	Krebs
2	28	70399	Lungenentzündung
3	35	84123	HIV
4	31	80112	Krebs
5	38	84156	HIV
6	34	80446	Lungenentzündung
7	36	70368	Krebs
8	33	70354	HIV
9	35	70491	Krebs

Abbildung 2.2: Beispieldaten ohne Anonymisierung

In Abbildung 2.2 sind beispielhaft erstellte Daten dargestellt. Vor der Anonymisierung sind die Daten durch die Spalte ID eindeutig zuordenbar. Das Alter und die Postleitzahl sind in diesem Fall Beispiele für Quasi-Identifier. Betrachtet man diese Attribute einzeln, lässt sich keine Verbindung zu einer Person herstellen. Zusammen mit dem Alter ist es aber ggfs. möglich, Personen eindeutig zu identifizieren. Das sensitive und damit zu schützende Attribut ist in diesem Beispiel die Krankheit.

Damit k-anonymity erreicht wird, wird zum einen die Spalte ID komplett entfernt. Zudem müssen die Quasi-Identifier so verallgemeinert werden, dass immer mindestens k Einträge mit gleichen Quasi-Identifier entstehen.

	Quasi-Identifier		**Sensitives Attribut**
	Alter	**PLZ**	**Krankheit**
C1	20 – 29	70*	Krebs
	20 – 29	70*	Lungenentzündung
C2	30 – 39	84*	HIV
	30 – 39	84*	HIV
C3	30 – 39	80*	Krebs
	30 – 39	80*	Lungenentzündung
C4	30 – 39	70*	Krebs
	30 – 39	70*	HIV
	30 – 39	70*	Krebs

Abbildung 2.3: k-anonyme Daten mit k=2 (2-anonymity)

Abbildung 2.3 zeigt die in Abbildung 2.2 erstellten Daten, nachdem die Quasi-Identifier so verallgemeinert wurden, dass es jeweils mindestens zwei (im letzten Fall, Cluster C4, sogar drei) gleiche Kombinationen der Quasi-Identifier gibt. Die Daten sind somit k-anonym.

In diesem Beispiel zeigt sich jedoch auch eine große Schwäche von k-anonymity. Zwar ist eine eindeutige Zuordnung eines Datensatzes zu einer Person nicht mehr möglich. Im Falle des Clusters C2 ist dies jedoch auch nicht nötig. Auf Grund des Wissens, dass eine Person in Cluster C2 ist, kann direkt auf das sensitive Attribut (in diesem Fall HIV) geschlossen werden, da in diesem Cluster alle Einträge denselben Attributwert haben.

Um das zu verhindern, wurde der Ansatz um das Prinzip der l-diversity erweitert [MKGV07]. Das Prinzip besagt, dass in jedem Cluster mindestens l „well-represented“, also gut repräsentierte Werte für die sensitiven Attribute existieren müssen. In [MKGV07] werden mehrere unterschiedliche Möglichkeiten geschildert, was „gut repräsentiert“ bedeutet. Die einfachste und gängigste Erklärung besagt, dass mindestens l verschiedene Werte vorhanden sein müssen. Damit lässt sich verhindern, dass auf Grund des Wissens, dass eine Person in einem Cluster abgebildet ist, automatisch auch das sensitive Attribut offengelegt wird. Auch bei diesem Prinzip gilt: Je höher l desto stärker die Anonymität.

Abbildung 2.4 zeigt den bereits bekannten Datensatz. Um l-diversity zu erreichen, wurde die Postleitzahl stärker verallgemeinert. Dadurch verschmelzen C2 und C3 zu einem Cluster. Nun sind in jedem Cluster mindestens zwei unterschiedliche sensitive Attribute vorhanden.

	Quasi-Identifier		**Sensitives Attribut**
	Alter	**PLZ**	**Krankheit**
C1	20 – 29	7*	Krebs
	20 – 29	7*	Lungenentzündung
C23	30 – 39	8*	HIV
	30 – 39	8*	HIV
	30 – 39	8*	Krebs
	30 – 39	8*	Lungenentzündung
C4	30 – 39	7*	Krebs
	30 – 39	7*	HIV
	30 – 39	7*	Krebs

Abbildung 2.4: l-diverse Daten mit l=2 (2-diversity). Außerdem wurde t-closeness (mit t=0,33) auf diese Daten angewendet

Ein drittes Konzept in diesem Zusammenhang nennt sich t-closeness [LLV07]. Es ist die Erweiterung von l-diversity und senkt das Risiko einer Re-Identifizierung weiter. Eine Schwachstelle beim l-diversity Ansatz ist die Verteilung der sensitiven Werte in einem Cluster. Ist eine Tabelle zum Beispiel 2-divers, so kann die Verteilung der sensitiven Werte in einem Cluster im schlechtesten Fall 9:1 zu sein. Damit ist zwar dann nicht eindeutig gesichert, dass die Personen, die dem Cluster zugeordnet werden können, einen bestimmten sensitiven Wert haben, die Wahrscheinlichkeit liegt aber immerhin bei 90%. Außerdem lassen sich ggfs. auch durch das Wissen über die Verteilung von sensitiven Werten weitere Rückschlüsse ziehen. Um dies zu verhindern, besagt das t-closeness Prinzip, dass die Verteilung der Werte der sensitiven Attribute innerhalb eines Clusters nicht zu stark von der Verteilung der sensitiven Werte des gesamten Datensatzes abweichen darf. Die Nähe der Clusterverteilung zur Verteilung im gesamten Datensatz wird mit t beschrieben. Die Beachtung dieses Prinzips verbessert die Anonymität weiter, reduziert aber noch stärker die Nutzbarkeit der Daten. In den Beispieldaten in Abbildung 2.4 wurde t mit Hilfe der in

[LLV07] beschriebenen „variational distance“ berechnet. Der größte Abstand zur gesamten Verteilung der sensitiven Attribute hat Cluster C1 mit 0,33. Je näher t an 0 liegt, desto gleicher ist die Verteilung der sensitiven Attribute in dem Cluster.

	Quasi-Identifier		Sensitives Attribut
	Alter	**PLZ**	**Krankheit**
C14	20-39	7*	Krebs
	20-39	7*	Lungenentzündung
	20-39	7*	Krebs
	20-39	7*	HIV
	20-39	7*	Krebs
C23	20-39	8*	HIV
	20-39	8*	HIV
	20-39	8*	Krebs
	20-39	8*	Lungenentzündung

Abbildung 2.5: Noch stärker verallgemeinerte Daten mit k=4, l=3 und t=0,19

Abbildung 2.5 zeigt ein Clustering mit stärkerem t. Jedoch gingen durch die nötige Verallgemeinerung des Alters noch mehr Daten verloren und die Datennutzbarkeit wurde weiter reduziert. In den Beispielen wird außerdem deutlich, dass die Art und Weise, wie die Cluster generiert werden, erheblichen Einfluss auf die benötigte Stärke der Generalisierung und damit verbunden die Nutzbarkeit der Daten hat.

2.2.2 Differential Privacy

Die Idee von Differential Privacy [Dwo08] unterscheidet sich grundlegend vom im vorherigen Abschnitt beschriebenen Vorgehen. Bei pertubationsbasierten Verfahren werden die Daten nicht verallgemeinert sondern durch Hinzufügen von Rauschen verändert. Dies passiert typischerweise, indem eine bestimmte Anzahl an sensitiven Attributen durch zufällige Werte ersetzt werden. Dadurch ist es im Nachhinein nicht mehr möglich zu erkennen, ob einzelne Werte wahr sind, oder zufällig erzeugt wurden. Wie oft sensitive Werte durch Zufallswerte ersetzt werden, steuert bei Differential Privacy der Parameter ϵ. Je geringer ϵ gewählt wird, desto stärker wird das Rauschen und damit die Anonymität

des Datensatzes. Mit Hilfe statistischer Verfahren lassen sich die mit Rauschen versetzten Daten trotzdem auswerten, wenn Informationen über die Stärke und die Art und Weise der Rausch-Generierung bekannt sind. Damit ist es nicht mehr möglich, Aussagen über einzelne Werte und damit persönliche Informationen zu treffen. Trotzdem können die anonymisierten Daten im Gesamten betrachtet und ausgewertet werden.

Durch mehrmalige gleiche Abfragen der Daten können jedoch Informationen über die Art und Weise der Rauschgenerierung gewonnen werden. Mit Hilfe dieser gewonnenen Informationen lässt sich nach einer ausreichenden Menge an Abfragen das Rauschen herausmitteln und dadurch die Daten de-anonymisieren. Um dies zu verhindern, wird in der Literatur von einem sogenannte Privacy Budget gesprochen, der Anzahl an Abfragen, die möglich sind, ohne dass zu viel Information über die Rauschgenerierung preisgegeben wird. Je geringer ϵ und damit je höher die Anonymisierung desto geringer ist das Privacy Budget und damit die Anzahl an möglichen Abfragen. Dieser Umstand bedeutet eine große Herausforderung in der Praxis, da die Anzahl der Abfragen bei ausreichender Anonymität oftmals nur noch sehr gering ist. Um das Verfahren sinnvoll anzuwenden sind deshalb sehr große Datenmengen nötig, um zumindest eine gewisse Anzahl an Abfragen auf die Daten zuzulassen.

Ein sehr großer Vorteil dieses Verfahrens ist die mathematisch klar definierbare Stärke der Anonymisierung über den Parameter ϵ. Wenn sich an das Privacy Budget gehalten wird, kann dieser Ansatz Anonymität von Daten garantieren, was so in anderen Verfahren nicht möglich ist. Die Anonymität der Daten ist in diesem Fall auch nicht von zusätzlichen Daten Dritter abhängig wie zum Beispiel bei k-anonymity.

3 Systematische Literaturrecherche

3.1 Vorgehen

Das ursprüngliche Ziel dieser Arbeit war es, aktuelle Ansätze und Implementierungen zur Anonymisierung von Daten im Automobilbereich zusammenzutragen und auszuwerten. Zur Suche nach Dokumenten und Ausarbeitungen wurde die Onlineplattform scopus[1] genutzt. Scopus indiziert und speichert wissenschaftliche Arbeiten aus verschiedensten Quellen, die mit Hilfe einer umfassenden Suche gefunden werden können. Zum Zeitpunkt des Schreibens dieser Arbeit waren auf scopus ca. 87 Millionen Dokumente verfügbar. Erste Suchergebnisse auf scopus zeigten jedoch schnell, dass zum aktuellen Zeitpunkt kaum Literatur zur Anonymisierung im Automobilbereich existiert. Die Suche nach „anonymization + automotive“ lieferte beispielsweise lediglich 8 Treffer. Auch die Suchen mit anderen Stichwörtern („anonymization + vehicle“, „anonymization + car“) lieferten kaum relevante Ergebnisse.

Die Literaturrecherche soll sich deshalb im Folgenden auf Anonymisierung im Allgemeinen beziehen und einen Überblick über aktuelle Konzepte, Ideen und Ansätze geben. Um eine möglichst objektive Darstellung der Ergebnisse zu gewährleisten, sollen im Folgenden alle Schritte von der Auswahl der grundsätzlich zu untersuchenden Dokumente bis zur Analyse der tatsächlich ausgewählten Arbeiten aufgeführt und begründet werden.

Die Basis für die Analyse der Arbeiten bietet die Dokumentensuche auf scopus. Mit Hilfe des Suchstrings „anonymization AND (approach OR technique OR algorithm) AND novel“ wurden 480 Arbeiten gefunden. Bei einer Suche nach „anonymization“ ist die Menge der Arbeiten mit rund 5200 Dokumenten zu groß. Bei den Ergebnissen sind oftmals irrelevante Arbeiten dabei, die mit der tatsächlichen Anonymisierung von Daten nichts zu tun haben. Dies liegt sicherlich auch an der im Allgemeinen oftmals falsch verwendeten Nutzung des Begriffs Anonymisierung. Um die Ergebnisse weiter einzuschränken und sich auf Ansätze und Algorithmen zu fokussieren, wurde die Suche mit Hilfe unterschiedlicher Stichwörter weiter verfeinert. Abbildung 3.1 zeigt die verschiedenen Suchstrings und die jeweilige Anzahl an Dokumenten, die scopus für diese Suche liefert.

Die mit dem zweiten und dritten Suchstring gefundenen Dokumente sind zwar relevanter, jedoch mit fast dreieinhalb Tausend Ergebnissen immer noch sehr groß. Zudem sind in den Arbeiten oftmals Vergleiche oder Implementierungen von bereits bestehenden

[1] https://www.scopus.com/

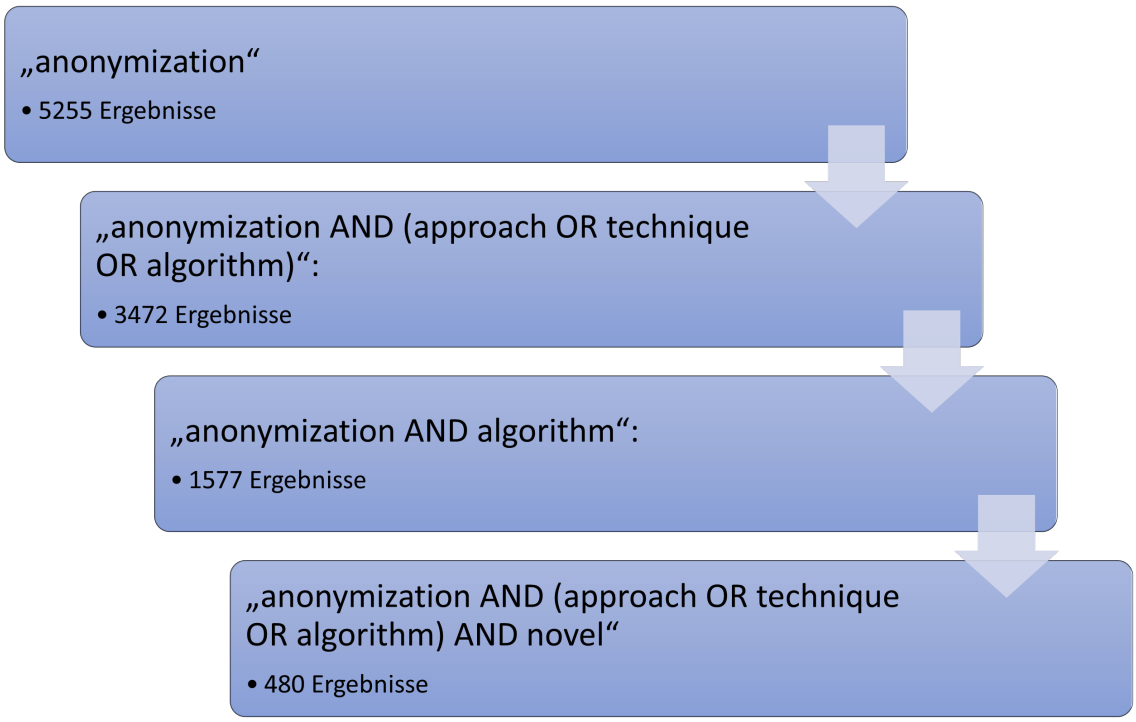

Abbildung 3.1: Anzahl der gefundenen Dokumente mit verschiedenen Suchstrings auf scopus

Ansätzen enthalten. Deshalb wurde die Suche um den Begriff „novel“ ergänzt, was zu den bereits genannten 480 Ergebnissen führt. Durch die Einschränkungen des Fachgebiets auf „Computer Science“, „Engineering“ oder „Mathematics“ sowie die Sprache Englisch wurden nochmals irrelevante Dokumente aussortiert, sodass am Ende 426 Dokumente übrig blieben, die tatsächlich in die Literaturrecherche einfließen.

Um die Dokumente strukturiert auszuwerten, wurde der PRISMA-Ansatz gewählt [PMB+21]. Der PRISMA-Ansatz ist ein Leitfaden für die Durchführung von Literaturrecherchen und bietet neben Checklisten einen klaren Ablauf zur Evaluierung der gefundenen Dokumente einer Recherche. Mit Hilfe klarer Ein- und Ausschlusskriterien sollen die gefundenen Arbeiten systematisch bearbeitet werden und sich die Ergebnismenge mit jedem Schritt reduzieren. Der PRISMA-Ansatz ist ein offener, iterativer Prozess, bei dem zu jeder Zeit neue Dokumente hinzugefügt und die verschiedenen Schritte neu evaluiert werden können.

Für die Auswertung der einzelnen Schritte wurden die Dokumente chronologisch absteigend analysiert, da in dieser Ausarbeitung die neuesten Erkenntnisse zur Anonymisierung betrachtet werden sollen. Ab 2015 sind bei der ersten Durchsicht der Ergebnisse keine nennenswert neuen Erkenntnisse mehr zu erwarten. Zudem sind die Arbeiten oder Ansätze teilweise bereits durch neuere Arbeiten überholt. Deshalb wurde die ausgiebige Analyse der Dokumente ab 2015 abgebrochen.

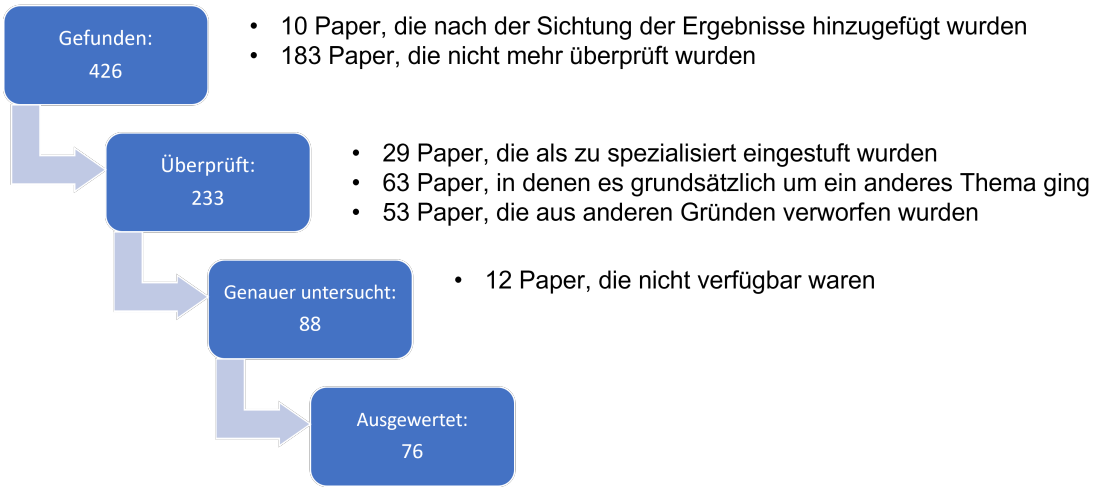

Abbildung 3.2: Einzelne Schritte bei der strukturierten Literaturrecherche nach dem PRISMA-Ansatz.

Abbildung 3.2 zeigt die einzelnen Schritte der Literaturrecherche, beginnend mit den 426 Dokumenten aus der oben bereits beschriebenen Suche auf Scopus. Von den 426 Ergebnissen sind 10 Dokumente erst nach der Sichtung der Ergebnisse hinzugefügt worden. Deshalb wurden diese Arbeiten nicht mehr berücksichtigt. Zudem wurden die 183 Dokumente ab 2015 nicht mehr näher ausgewertet. Von den übrigen 233 Arbeiten wurden 29 als zu spezialisiert eingestuft. In diesen Arbeiten geht es um spezielle Einzelfälle, die für eine Auswertung in diesem Rahmen keine Erkenntnisse bringen (beispielsweise [KLKH22]. Zudem wurden 63 Arbeiten verworfen, in denen es grundsätzlich um ein anderes Thema ging und Anonymisierung in einem anderen Zusammenhang erwähnt wurde. 53 Paper wurden aus anderen Gründen verworfen. Hierunter fallen beispielsweise Arbeiten, die sich mit Angriffen bzw. De-Anonymisierung beschäftigen. Außerdem wurden Arbeiten verworfen, die lediglich bereits bestehende Ansätze vergleichen, aber keine neuen Ansätze oder Ideen liefern. Des Weiteren wurde in einige Arbeiten die Implementierung von bereits bekannten Anonymisierungsverfahren in Umgebungen oder Frameworks behandelt. Auch diese Arbeiten wurden nicht weiterverfolgt. Von den restlichen 88 Arbeiten waren 12 Paper nicht zugänglich oder verfügbar. Damit bleiben 76 Arbeiten, die die Grundlage für die im folgenden Abschnitt erläuterten Ergebnisse bilden.

3.2 Ergebnisse

Zur Auswertung der Veröffentlichungen wurden alle analysierten Paper zunächst kategorisiert. Für jede Arbeit wurde der grundsätzliche Ansatz und die damit verbundene Methodik festgelegt. Einige Arbeiten nutzen Ansätze, die in mehrere Kategorien einzuordnen sind. So wird beispielsweise in [AAS+18] eine Kombination aus Verschlüsselung (Blockchain) und Differential Privacy angewendet. In diesen Fällen wurde die Arbeit der Methodik

zugeordnet, die den größten Anteil in der Ausarbeitung darstellt. Des Weiteren wurden die Paper einem Anwendungsbereich zugeordnet, soweit dies möglich war. In Abschnitt 3.2.2 wird näher erläutert, in welche Anwendungsbereiche die Ergebnisse in der Literaturrecherche eingeteilt wurden, sowie deren Charakteristiken. Zuletzt wurden die Ausarbeitungen nach Datentyp kategorisiert. Der Datentyp beschreibt in diesem Fall die Art der Daten, die anonymisiert werden sollen.

Mit Hilfe dieser Kategorisierungen soll die vorhandene Literatur analysiert und die wichtigsten Erkenntnisse dieser Literaturrecherche erläutert werden. Einzelne Arbeiten werden teilweise als Beispiel referenziert. Es sollen jedoch nicht alle Arbeiten detailliert aufgelistet und analysiert werden. In Anhang A sind alle analysierten Arbeiten aufgeführt.

Unabhängig von den oben aufgeführten Kategorisierungen wurden zudem die verschiedenen Methoden zur Evaluierung der Ansätze betrachtet. Abschnitt 3.2.4 gibt eine Übersicht, wie neue Ansätze mit bisherigen Methoden verglichen werden, welche Metriken genutzt werden und auf welche verschiedenen Aspekte eingegangen wird.

3.2.1 Methodiken

Zunächst soll ein Blick auf die verschiedenen Methodiken geworfen werden, die im Rahmen dieser Literaturrecherche analysiert wurde. Abbildung 3.3 zeigt die Verteilung der verschiedenen Methodiken in den analysierten Arbeiten. Mit fast 70% haben die aggregationsbasierten Ansätze mit Abstand die größte Verwendung. Hierbei handelt es sich in den meisten Fällen um Implementierungen von k-anonymity. Bei nahezu allen Papern werden die erweiterten Konzepte l-diversity und t-closeness ebenfalls berücksichtigt. Im Großen und Ganzen können drei Arten von Arbeiten in diesem Bereich unterschieden werden:

- Clustering-Algorithmen und Verfahren zur besseren Hierarchisierung
- Anwendung von k-anonymity auf andere Daten oder Datentypen
- Erweiterung von k-anonymity, l-diversity und t-closeness um neue Konzepte zur Steigerung der Datennutzbarkeit und Anonymisierung

Die Erzeugung und Generierung der Cluster bei k-anonymity hat große Auswirkungen auf die Nutzbarkeit der Daten. So müssen bei einem guten Cluster-Verfahren die Attribute weniger stark generalisiert werden um k gleiche Einträge derselben Quasi-Identifier zu erzeugen. Folglich gibt es immer wieder neue Ansätze, Cluster noch effizienter zu erstellen bzw. zu berechnen, um damit die Nutzbarkeit weiter zu steigern ([SVSJ19], [BJW18], [GYG+19], [WNV18], [ZGB+18], [SNNR22], [KZD22]). Auch spielt die Art der Hierarchisierung, also welche Daten wie generalisiert werden, teilweise eine große Rolle. Neue Ideen wie vor allem bei Positionsdaten Attribute generalisiert werden können, finden sich zum Beispiel in [VDW+21] oder [HRS20].

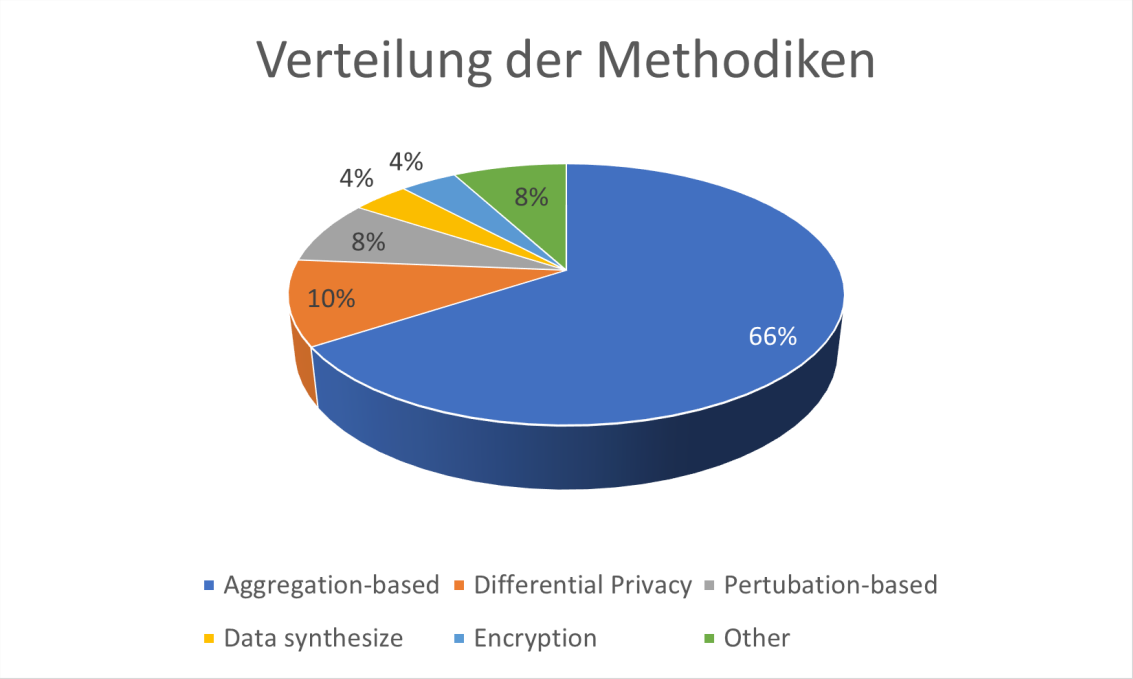

Abbildung 3.3: Verteilung der verschiedenen Methodiken in den analysierten Arbeiten.

Das Konzept von k-anonymity basiert wie bereits in Abschnitt 2.2.1 beschrieben unter anderem auf der Unterscheidung der Attribute in Quasi-Identifier und sensitive Attribute bei tabellarischen Daten. Es gibt jedoch Anwendungsfälle, bei denen es sich entweder um eine andere Art von Daten handelt (siehe Abschnitt 3.2.3) oder bei denen eine eindeutige Zuordnung zu Quasi-Identifiern und sensitiven Attributen nicht ohne weiteres möglich ist. [SOTO19] beschreibt beispielsweise die Anwendung von k-anonymity, l-diversity und t-closeness auf Daten, die sensitive Quasi-Identifier enthalten. [RN22] beschreibt einen neuen Ansatz für Daten, die mehrere sensitive Attribute enthalten. Auf die Anpassung von k-anonymity für verschiedene, andere Arten von Daten wird in Abschnitt 3.2.3 näher eingegangen. Außerdem werden in Abschnitt 3.2.2 nochmals verschiedene aggregationsbasierte Verfahren näher beleuchtet.

Zuletzt gibt es einige Arbeiten, die sich mit der noch stärkeren Anonymisierung aggregationsbasierter Ansätze beschäftigen. Dazu werden weitere Parameter eingeführt, die die Re-Identifikationswahrscheinlichkeit noch stärker reduzieren ([KTA+22], [LGZ+18]).

Der nächst größte Anteil mit rund 10% haben Ansätze, die Differential Privacy nutzen. Differential Privacy kann den pertubationsbasierten Verfahren zugeordnet werden. Da Differential Privacy jedoch noch weitere Merkmale hat und als Ansatz auch eigenständig in vielen Arbeiten erscheint, wurde bei der Einordnung der Methodik ebenfalls zwischen Differential Privacy und pertubationsbasierte Verfahren unterschieden (gemeinsam ca. 20% der Arbeiten).

Die eine Hälfte der Arbeiten, die Differential Privacy nutzen, beschäftigen sich mit der Verbesserung der Nutzbarkeit der Daten durch Anpassung der mathematischen Verfahren oder durch zusätzliche Schritte, in denen die Daten vorbereitet oder angepasst werden ([JHRZ22], [Ham17]. [Alj21], [GL19c]. Hierbei handelt es sich meistens um neue Verfahren zur Generierung des Rauschens. Der andere Teil nutzt Ansätze von Differential Privacy und wendet sie auf andere Arten von Daten an. So beschreibt [SG22] die Nutzung von Differential Privacy in Cloud Umgebungen, [LZX22] passt die Implementierung von Differential Privacy an Routenverläufe an und in [GL19a] wird neben einer Metrik zur Messung der Datennutzbarkeit Differential Privacy auf Graphen angewendet.

Die restlichen gut 10% der Arbeiten beschäftigen sich mit Verschlüsselung, synthetischer Datengenerierung und einzelnen weiteren Verfahren. Verschlüsselung ist wie bereits in Abschnitt 2.1 erläutert streng genommen kein Anonymisierungsverfahren. Trotzdem gibt es Ansätze, die mit Hilfe von Verschlüsselung, teilweise in Kombination mit anderen Verfahren, Daten verfremden und als Anonymisierung einstufen. Die hier aufgeführten Arbeiten nutzen dazu meist Blockchain-Verfahren ([JAM+21], [AAS+18]). Auf Blockhain-Verfahren soll in dieser Arbeit jedoch nicht näher eingegangen.

Synthetische Datengenerierung hat vor allem durch die Entwicklungen im Bereich des maschinellen Lernens in den vergangenen Jahren an Bedeutung gewonnen. Zur Generierung solch künstlicher Daten werden vermehrt Generative Adversarial Networks (GAN) eingesetzt. GAN bestehen aus zwei neuronalen Netzen. Ein Netz versucht, künstliche Daten auf Basis der tatsächlich existierenden Daten zu erzeugen, die die gleichen Charakteristiken aufweisen. Das andere neuronale Netz versucht zu erkennen, ob die vorgelegten Daten künstlich oder echt sind. Durch das Zusammenspiel der beiden Netze werden die künstlich erzeugten Daten immer besser, bis sie irgendwann kaum mehr von den echten Daten unterscheidbar sind. Diese künstlich generierten Daten können nun zusammen mit den echten Daten ausgewertet werden. Durch die gleiche Charakteristik der Daten bleibt die Nutzbarkeit erhalten. Es ist jedoch nicht mehr möglich zu sagen, welche Daten von echten Personen stammen und welche Daten künstlich erzeugt wurden. Dadurch kann Anonymität gewährleistet werden. In [PA20a] und [PA20b] werden beispielsweise solche Verfahren beschrieben.

3.2.2 Anwendungsbereich

Bei den analysierten Arbeiten wurde neben der grundsätzlichen Methodik auch der Anwendungsbereich festgelegt, soweit dieser ersichtlich war. Eine Übersicht der Anwendungsbereiche sowie der Anteil der Methodiken, die in dem jeweiligen Anwendungsgebiet genutzt werden, zeigt Abbildung 3.4.

Mit rund 40% ist der größte Teil der Paper nicht einem speziellen Anwendungsbereich zuordenbar. In diesen Arbeiten werden meistens theoretische Aspekte der jeweiligen Methodiken bearbeitet. So werden bei aggregationsbasierten Methoden ohne Anwendungs-

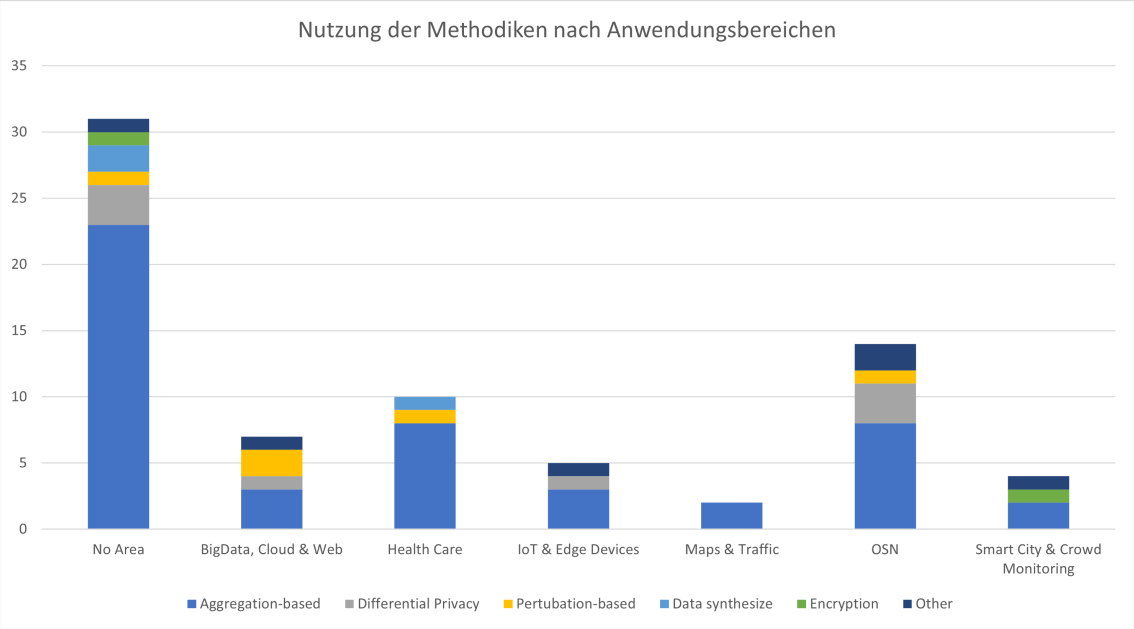

Abbildung 3.4: Anwendungsgebiete der analysierten Arbeiten mit der jeweiligen Verteilung der Methodiken in diesem Gebiet.

gebiet oft neue Clustering-Algorithmen vorgestellt oder neue Verfahren zur Steigerung der Anonymität (siehe Abschnitt 3.2.1). Auch für anderen Methodiken werden meistens neue mathematische Verfahren vorgestellt, die bereits bekannte Ansätze optimieren sollen.

Die zwei größten Anwendungsgebiete sind die Sozialen Netzwerke (Online Social Networks (OSN)) mit rund 20% und der Gesundsheitsbereich mit ca. 15%. Der Gesundheitsbereich hat sich durch die Auswertung klinischer Studien bereits früh mit Anonymisierungskonzepten beschäftigt. Nach wie vor gibt es einige Arbeiten, die die Anonymisierungsverfahren in diesem Bereich optimieren (z.B. [AM22], [ARPL21]). Anonymisierung in Sozialen Netzwerken ist ebenfalls ein großes Forschungsgebiet, da soziale Netzwerke immer mehr persönliche Daten enthalten, an denen eine Großzahl an Branchen für Analysen interessiert ist. Daten aus Sozialen Netzwerken sind oftmals sogenannte Graph Daten, also Knoten-Kanten-Modelle. Dadurch lässt sich der relative hohe Anteil an Differential Privacy ([GL19a], [GL19c]), pertubationsbasierten und anderen Verfahren ([ATA18], [RMTC18]) erklären (siehe auch Abschnitt 3.2.3). In allen anderen Bereichen - mit Ausnahme von BigData, Cloud & Web - sind die aggregationsbasierten Verfahren mit Abstand die am meisten verwendeten. Für den Bereich BigData, Cloud & Web lässt sich dies möglicherweise dadurch erklären, dass die zu anonymisierende Datenmenge im Vergleich zu anderen Bereichen sehr hoch ist, und sich somit pertubationsbasierte Verfahren ([EAZS18], [KBS21]) leichter einsetzen lassen.

3.2.3 Datentyp

Zuletzt wurden die analysierten Paper bezüglich der zu anonymisierenden Daten näher betrachtet. Abbildung 3.5 zeigt den Vergleich verschiedener Datentypen und erneut die Unterteilung in die jeweils genutzten Methodiken.

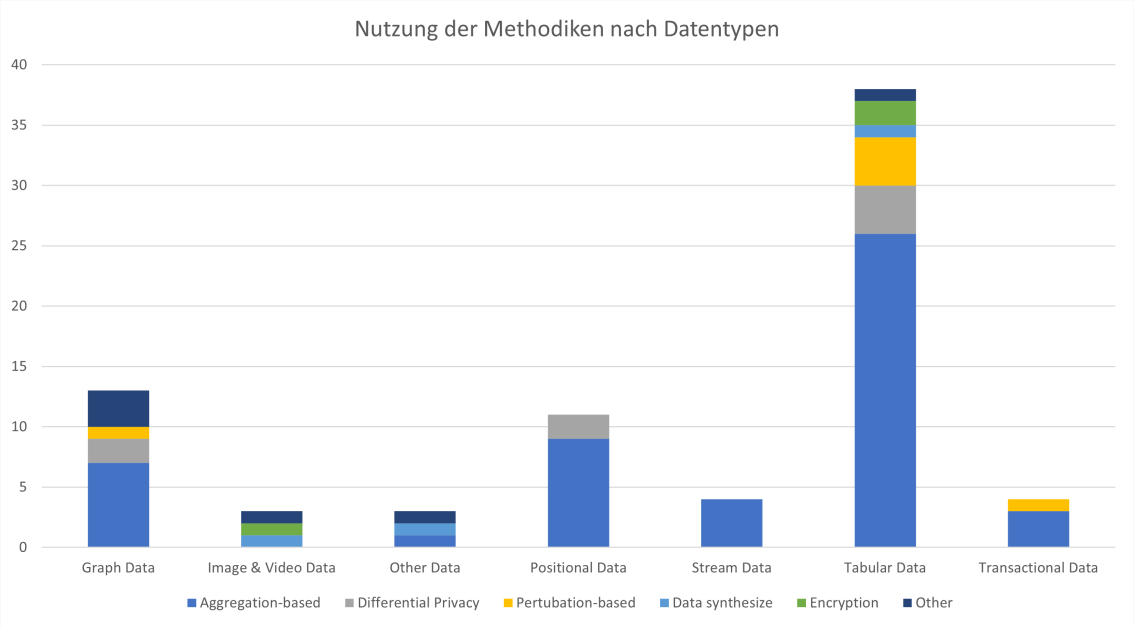

Abbildung 3.5: Datentypen der analysierten Arbeiten mit der jeweiligen Verteilung der Methodiken in diesem Gebiet.

Rund die Hälfte der Arbeiten beschäftigt sich mit tabellarischen Daten. Da ein Großteil der zu anonymisierenden Daten im Allgemeinen als tabellarische Daten vorliegt und die Anonymisierung bzw. Erarbeitung neuer Konzepte mit diesen Daten am einfachsten ist, ist dieser große Anteil nicht weiter verwunderlich. Auch ist die Verteilung der genutzten Methodiken bei tabellarischen Daten nahezu gleich, wie die Verteilung der Methodiken über alle Datentypen hinweg. Einige Arbeiten beschäftigen sich jedoch auch mit der Anonymisierung von Graph Data und Positional Data (jeweils rund 15%). Wie bereits in Abschnitt 3.2.2 beschrieben, kommen Graph Data (Knoten-Kanten-Modelle) zu einem sehr großen Anteil in Sozialen Netzwerken zum Einsatz. Es gibt jedoch vereinzelt auch andere Anwendungsgebiete. Bei diesen Daten ist der Anteil anderer Methodiken (neben aggregationsbasierten Verfahren) sehr hoch. Da Knoten-Kanten-Modelle sich stark von tabellarischen Daten unterscheiden, wurden einige andere Verfahren speziell für diesen Datentyp entwickelt ([TCC20], [GL19b]), die auf der Veränderung der Knoten-Kanten-Modelle basiert und oftmals nicht aggregationsbasierten Verfahren zuordenbar sind.

Der zweite große Bereich, Positionsdaten, lässt sich nochmals in zwei Unterarten von Daten aufteilen. Zum einen gibt es Arbeiten, die sich mit dem Anonymisieren von einzelnen Orten, sogenannten Point-Of-Interest (POI), beschäftigen ([ALWJ18], [LZW21],

[SO17]). Zum anderen werden in einigen Arbeiten Routendaten (engl. Trajectories) anonymisiert ([WLM17], [MDM22], [LZX22]). Bei Positionsdaten werden hauptsächlich aggregationsbasierte Verfahren angewendet. Für diese Daten ist eine Anonymisierung jedoch deutlich komplexer als für tabellarische Daten, da Positionsdaten teilweise stark verändert oder verallgemeinert werden müssen, um keine Rückschlüsse mehr auf einzelne Personen zuzulassen. Zudem wird meistens eine ausreichend große Menge an Daten benötigt, um die Verfahren sinnvoll einzusetzen.

Ansätze für Streamingdaten, also zusammenhängende Datenströme, sowie für Transaktionale Daten wurden in den ausgewerteten Arbeiten nur vereinzelt gefunden. Für diese Datentypen wurden typischerweise bereits existierende Ansätze genutzt und diese auf die Charakteristik des jeweiligen Datentyps angewendet ([MGN20], [TWTH20], [PKS22]). Arbeiten, die sich mit Bild - und Videodaten beschäftigen, wurden ebenfalls kaum analysiert. Diese Daten unterscheiden sich in ihrer Art und auch der Form, wie sie anonymisiert werden müssen, sehr stark von den bisher beschriebenen Datentypen. Folglich werden dafür auch andere Verfahren eingesetzt. Evaluierung von Ansätzen zur Bild- und Videodaten Anonymisierung genauso wie Arbeiten zur Anonymisierung von Audiodaten waren jedoch nicht im Fokus dieser Literaturrecherche.

3.2.4 Evaluierung der Methoden

Neben der Einordnung in die zuvor aufgezeigten Kategorien soll in diesem Abschnitt auf die Evaluierung der Ansätze in den Papern eingegangen werden. Hierbei soll vor allem beleuchtet werden, wie neue Ansätze mit bisher existierenden Ansätzen verglichen werden, welche Eigenschaften einfließen und mit welchen Metriken diese gemessen werden. Es wurden jedoch nicht wie in den Abschnitten zuvor alle Arbeiten kategorisiert, sondern stichprobenartig einige Arbeiten genauer angeschaut und auf die beschriebene Thematik untersucht. Anschließend wurden die beschriebenen Ergebnisse durch weitere Paper verifiziert.

Im Großen und Ganzen können drei Eigenschaften identifiziert werden, die beim Vergleich von Ansätzen herangezogen werden:

- Die Datennutzbarkeit
- Die Ausführungsdauer (engl. Execution Time)
- Die (Stärke der) Anonymität

Datennutzbarkeit

In fast allen untersuchten Arbeiten wird die Datennutzbarkeit im Vergleich zu anderen Ansätzen gemessen und verglichen. Dazu wird meistens vom sogenannten Information Loss gesprochen, also der Menge an Information, die durch den Vorgang der Anonymisierung

verloren geht. Wie der Information Loss definiert wird, variiert jedoch stark von Paper zu Paper. So wird der Information Loss vor allem bei aggregationsbasierten Methoden über die Stärke der Generalisierung der Quasi-Identifier definiert. Unterschieden wird immer zwischen kategorischen und numerischen Attributen. Die genaue Definition unterscheidet sich jedoch meistens leicht voneinander (vgl. [LGZ+18], [VDW+21], [KSC14]). In einigen Paper wird die sogenannte Normalized Certainty Penalty (NCP) verwendet um Ansätze zu vergleichen ([MGRM19], [GLY+17], [WL18]). Bei Anwendungsfällen, in denen die zu anonymisierenden Daten anschließend kategorisiert, klassifiziert oder geclustert werden, wird oftmals der Klassifikationsfehler oder die Accuracy (Genauigkeit) zur Messung des Information Loss genutzt. Dazu wird der Classification Error bzw. die Accuracy vor der Anonymisierung und nach der Anonymisierung verglichen ([HYL+17], [SVSJ19], [BJW18]).

Zuletzt gibt es auch einige Paper ([EM19], [KP19], [SOTO19], [JHRZ22], [Alj21]), die andere Metriken definieren, mit denen anschließend die Ergebnisse der Anonymisierung verglichen werden. Vor allem bei Ansätzen, die nicht den aggregationsbasierten Methoden zuzuordnen sind, sind individuelle Metriken zur Messung der Datennutzbarkeit zu finden.

Execution Time

In sehr vielen Arbeiten wird die Execution Time des vorgestellten Ansatzes mit anderen Ansätzen verglichen. Die Execution Time spielt zwar in den meisten Fällen keine große Rolle, jedoch ist dieser Wert sehr leicht zu erfassen und muss zumindest bei sehr großen Datenmengen oder bei Echtzeitdaten (siehe Live-Traffic in Abschnitt 4.1) berücksichtigt werden. Zudem lassen sich durch die Execution Time Rückschlüsse auf die Komplexität des Algorithmus ziehen.

Anonymität

Betrachtet man in den analysierten Arbeiten den Vergleich der vorgestellten Arbeiten zu bisher existierenden Ansätzen hinsichtlich Anonymität, zeigen sich einige interessante Erkenntnisse. So gibt es zahlreiche Paper, in denen auf die Stärke der Anonymität gar nicht eingegangen wird ([LGZ+18], [], [Alj21]). Dies sind oft aggregationsbasierte Verfahren, in denen die Anonymität über die Parameter k,l,t usw. definiert wird. In den Ansätzen geht es oftmals um die Steigerung der Datennutzbarkeit. Da die Einschränkungen der Cluster bezüglich der aufgezählten Anonymisierungsparameter eingehalten werden, wird bei der Evaluierung der Ergebnisse offensichtlich nicht mehr auf die Stärke des Verfahrens hinsichtlich Anonymisierung eingegangen.

Des Weiteren gibt es einige Arbeiten, die den vorgestellten Ansatz bezüglich Stärke der Anonymisierung lediglich als Fließtext diskutieren ([KP19], [ZLL17], [GYG+19]). Hierbei fehlt zumeist ein formaler, mathematischer Vergleich der Stärke von Anonymität.

In den Arbeiten, in denen die Stärke der Anonymisierung in einer mathematischen Weise mit anderen Arbeiten verglichen wird ([EAZS18], [JHRZ22], [VDW+21], [WL18]), werden diese Metriken meistens speziell definiert. Hierbei gibt es zum einen Metriken, die im Paper selbst, teilweise als Teil der Arbeit, definiert und erläutert werden. Oftmals wird jedoch auf Metriken verwiesen, die in einem anderen Paper definiert wurden. Auffällig ist, dass es kaum Arbeiten gibt, die dieselben Metriken zum Vergleich der Anonymität nutzen.

4 Transfer zur Automobilwelt

Nachdem im vorherigen Kapitel die Ergebnisse der Literaturrecherche erläutert wurden, soll nun eine Brücke zum Automobilbereich geschlagen werden. Ziel ist es, die allgemeinen Erkenntnisse der Recherche auf konkrete Beispiele bei der Anonymisierung von Fahrzeugdaten anzuwenden. In den folgenden Abschnitten soll es um die Anonymisierung von Daten gehen, die innerhalb des Autos anfallen. Nicht eingegangen wird auf andere Daten, die bei Autokonzernen in anderen Bereichen entstehen. Nichtsdestotrotz wird am Ende des Kapitels kurz auf grundsätzliche Probleme bzw. Herausforderungen im Automobilbereich eingegangen, in denen solche Aspekte mit einfließen.

4.1 Anwendungsfälle in der Automobilwelt

Wie bereits beschrieben, fallen in modernen Autos inzwischen sehr große Datenmengen an, die in unterschiedlichen Bereichen erfasst und gesammelt werden. Diese sollen für zahlreiche unterschiedliche Aufgaben genutzt werden. [Fie23] beschreibt unterschiedliche Anwendungsfälle bei der Erhebung von Fahrzeugdaten, in denen Anonymisierung eine Rolle spielt. Diese werden außerdem näher definiert und bewertet. Auf diese Anwendungsfälle soll im Folgenden näher eingegangen werden.

4.1.1 Product improvement

Für Produktverbesserungen werden bereits heute zahlreiche Daten in Autos aufgezeichnet. Diese können vom Hersteller ausgewertet werden, um somit weitere Erkenntnisse über bestimmte Vorgänge im Auto zu erlangen und dadurch verschiedene Komponenten zu verbessern. Die Anzahl an Beispielen für solche Verbesserungen ist groß: Von Verbesserung von Bremsen auf Grund des Fahrhaltens über effizienteren Benzinverbrauch bis hin zu Batterieverbesserungen für elektrische Autos, die in [Fie23] als Beispiel herangezogen werden. Die auf den ersten Blick harmlos erscheinenden Daten können jedoch unter Umständen trotzdem gewisse Rückschlüsse auf den Fahrer zulassen und müssen zumindest rudimentär anonymisiert werden. Wenn beispielsweise für die Verbesserung von Batterien erhoben wird, an welchen Ladesäulen das Auto geladen wurde, können über den Standort der Ladesäule auch sensiblere Benutzerdaten erhoben und verarbeitet werden. Die Daten, die in diesem Anwendungsfall anfallen, sind größtenteils tabellarische Daten. Da die Daten keine Verhaltensdaten von Nutzern beinhalten, ist eine geringere Anonymität

vermutlich in den meisten Fällen ausreichend. Zudem lassen sich im Falle einer Re-Identifizierung kaum kritische Erkenntnisse über den Dateninhaber gewinnen. Abhängig von der nötigen Genauigkeit der Daten für anschließend stattfindende Auswertungen dürften aggregationsbasierte Ansätze mit relativ kleinen Werten für k, l oder t bei den entsprechenden Konzepten vielversprechende Ansätze für diesen Bereich sein. Geringe Anonymität lässt eine höhere Datennutzbarkeit durch geringe Generalisierung zu. Durch die vermutlich sehr hohe Anzahl an Daten und der bereits beschriebenen niedrigeren Notwendigkeit von Anonymität sind vermutlich auch Ansätze mit Differential Privacy gut möglich. Mit ausreichend hoher Datennutzbarkeit und dadurch geringerem Hinzufügen von Rauschen dürfte dieser Ansatz noch genauere Ergebnisse liefern. Letztlich wird die genaue Art der Daten, die für die Auswertungen benötigt werden entscheiden, welcher Ansatz in diesem Fall am sinnvollsten einzusetzen ist.

4.1.2 Live Traffic

Der zweite Anwendungsfall beschreibt die Erfassung und Nutzung der aktuellen Verkehrslage. Die meisten Navigationssysteme berechnen Routen heutzutage auch auf Basis aktueller Verkehrslagen, um beispielsweise Staus oder Baustellen zu umfahren. Die Daten für die Berechnung von Verkehrslagen können über verschiedene Sensoren im Auto erhoben werden. Neben GPS-Daten, verbunden mit dem Zeitstempel, werden auch Geschwindigkeit, Abstand zu anderen Autos und weitere Daten erfasst. Bei diesen Daten handelt es sich um hoch sensible, persönliche Daten, bei denen eine starke Anonymisierung unabdingbar ist. Zur Anonymisierung der GPS-Daten muss unterschieden werden, ob lediglich einzelne Orte anonymisiert werden müssen (POI) oder gesamte Routen (Trajectories). Für beide Datentypen gibt es wie in Abschnitt 3.2.3 näher beschrieben verschiedene Ansätze, die zu vielversprechenden Ergebnissen führen. Letztlich werden jedoch auch einige erhobene Daten verworfen werden müssen, da in sehr entlegenen Gegenden eine Anonymisierung von GPS-Daten auf Grund fehlender weiterer Quellen nicht möglich sein dürfte. Für die anderen erhobenen Daten (Geschwindigkeit, Nähe zu anderen Autos) können wie bereits im vorherigen Abschnitt beschrieben sowohl aggregationsbasierte als auch pertubationsbasierte Verfahren angewendet werden. Geprüft werden müsste, ob die anfallenden Daten als einzelne Datenpunkte anonymisiert und im Folgenden verwendet werden können, oder ob die Daten - beispielsweise durch die semantische Verknüpfung von GPS-Daten - als zusammenhängende Sequenz analysiert werden sollen. In diesem Fall wären das vermutlich transaktionale Daten, für die entsprechend angepasste Anonymisierungsmethoden verwendet werden müssten.

Ein weiterer wichtiger Punkt bei diesem Anwendungsfall ist die Dauer und auch der Ort der Anonymisierung. Daten über die aktuelle Verkehrslage sind Echtzeitdaten, die möglichst schnell nach Erfassung verwendet werden müssen. Werden zum Beispiel Staus identifiziert, müssen diese möglichst schnell den Navigationssystemen zur Verfügung gestellt werden, wenn diese mit Hilfe dieser Information Routenänderungen berechnen sollen. Damit spielen beim eingesetzten Anonymisierungsverfahren die Dauer und Komplexität der

Anonymisierung eine wesentliche Rolle. Des Weiteren muss geprüft werden, ob eine zentrale Anonymisierung in Intervallen nach Erfassung der Daten von mehreren Fahrzeugen möglich ist, ob dafür genügend Daten vorhanden sind, oder ob ggfs. eine Anonymisierung an Bord des Autos nötig ist. In der Literatur wird vereinzelt zwischen Static Data und Dynamic Data [PA20b] unterschieden. Hierbei geht es jedoch meistens um die Unterscheidung, ob es sich um abgeschlossene Daten handelt, zu denen keine weiteren Datensätze mehr hinzugefügt werden, oder ob die Daten kontinuierlich durch neue Daten ergänzt werden. Im vorliegenden Fall ist es jedoch womöglich sogar notwendig, veraltete Daten wieder zu entfernen. Ein Ansatz für diese Art Anwendungsfall ist in der in dieser Arbeit erstellten Literaturrecherche nicht enthalten.

4.1.3 Map Updating

Ein ähnlicher, wenn auch nicht gleicher Anwendungsfall sind Karten-Updates. Um Navigationssysteme auf dem neuesten Stand zu halten, müssen Karten regelmäßig um neue Informationen ergänzt werden. Beispiele hierfür wären Änderungen der Verkehrsführung (Kreisverkehr, Ampel, Einbahnstraßen) oder auch Änderungen von Tempolimits. Auch diese Informationen lassen sich am einfachsten über Autos gewinnen, die die entsprechenden Routen befahren. Ähnlich wie beim Anwendungsfall Live Traffic fallen auch hier GPS-Daten an. Informativ können aber auch noch weitere Daten über die Route sein, wie zum Beispiel Straßenschilder, die Fahrtgeschwindigkeit oder Aussagen über den Zustand der befahrenen Straße. Für die GPS-Daten gelten die gleichen Einschränkungen wie im vorherigen Anwendungsfall. Je nach Frequenz der erhobenen und verwendeten Daten können jedoch eher POI-Daten anfallen als gesamte Routen. Für die weiteren Daten gelten die bereits geschilderten Anmerkungen. Ein neuer und kritischer Punkt in diesem Anwendungsfall sind Kameradaten, die beispielsweise Straßenschilder aufnehmen und verwerten. Zum einen können auch auf Grund dieser Daten Rückschlüsse auf die Positionen des Autos gezogen werden. Zum anderen bergen diese Aufnahmen die Gefahr, dass andere Personen oder Autos auf den Kamerabildern sind. Diese Daten müssen entsprechend ebenfalls stark anonymisiert werden, um nicht sensitive Daten von Dritten ungeschützt zu verwenden. Auf Video- und Bildanonymisierung wird in dieser Arbeit jedoch nicht näher eingegangen.

4.1.4 Autonomous driving

Ein weiterer wichtiger Anwendungsfall ist die Datenerhebung für autonomes Fahren. Um die Entwicklung im Bereich autonomes Fahren voranzubringen, müssen echte Daten aufgezeichnet und den entsprechend eingesetzten Systeme zum Lernen zur Verfügung gestellt werden. Nur mit einer ausreichend großen Menge an Daten aus realen Bedingungen wird sich das autonome Fahren weiterentwickeln. Auch hier fallen verschiedenste sensitive Daten an. Neben Daten zum Fahrverhalten werden auch Kamera- oder andere Sensordaten (beispielsweise Radersensoren) aufgezeichnet sowie sämtliche weitere Daten, die Aussagen

über die Umgebung zulassen. Die große Menge an kontinuierlich anfallenden Daten könnte möglicherweise mit Hilfe von Differential Privacy Ansätzen anonymisiert werden. Die Menge der Daten dürfte für diesen Fall groß genug sein. Da es sich oft jedoch um ganze Sequenzen handelt, müssen auch Ansätze zur Anonymisierung transaktionaler Daten betrachtet werden. Je nach Implementation könnten sogar Ansätze zum Anonymisieren von Streamdaten erforderlich sein. Zuletzt sind für diesen Fall sicherlich auch Ansätze mit KI-basierten Methoden bzw. künstlich generierten Daten interessant.

4.1.5 Speech recognition

Im letzten Anwendungsfall geht es um die Datenerhebung von Sprachassistenten, die eine immer größere Rolle in modernen Autos spielen. Die Erfassung von Daten zur Verbesserung dieser Systeme ist daher naheliegend. Allerdings ist es in dieser Arbeit nicht das Hauptthema, die Anonymisierung von Sprache zu behandeln. Dennoch ist es wichtig zu betonen, dass der Schutz der Privatsphäre und die Anonymität der Nutzer auch in diesem Bereich eine große Bedeutung hat. Hierbei können die Anonymisierungsverfahren, die in anderen Domänen verwendet werden, voraussichtlich problemlos angewandt werden, da sich der Anwendungsfall kaum von den Sprachassistenten in anderen Bereichen unterscheidet. Es ist jedoch notwendig, diese Verfahren auf ihre Effektivität und Praktikabilität zu überprüfen, um sicherzustellen, dass die Anonymität der Nutzer gewahrt bleibt. Insgesamt ist die Erfassung von Sprachdaten ein wichtiger Schritt zur Verbesserung von Sprachassistenten und letztendlich auch der Fahrzeugtechnologie im Allgemeinen.

4.2 Weitere Herausforderungen

Neben den bereits aufgeführten Anwendungsfällen gibt es auch noch weitere Herausforderungen und Gebiete, die bei der Anonymisierung im Automobilbereich eine Rolle spielen. So muss beispielsweise geklärt werden, ob die Anonymisierung im Auto direkt oder anschließend im Rechenzentrum stattfindet. Muss die Anonymisierung im Fahrzeug stattfinden, so spielt wie bereits im Use-Case Live-Traffic skizziert die Execution Time eine große Rolle. Des Weiteren müssen Verfahren gefunden werden, wie einzelne Datensätze anonymisiert werden können. Möglicherweise können in diesem Fall Ansätze wie Federated Machine Learning (Föderiertes Maschinelles Lernen) genutzt werden. Beim Federated Machine Learning werden dezentral (in diesem Fall in den Fahrzeugen) lokale KI-Modelle mit Hilfe der erhobenen Daten trainiert. Die daraus entstehenden KI-Modelle gelten als anonymisiert, da keine Rückschlüsse auf einzelne Personen oder Datensätze gezogen werden können. Anschließend werden die dezentral, lokal trainierten KI-Modelle zu einem globalen KI-Modell zusammengeführt, das die Informationen aller lokalen Modelle beinhaltet ([KH22]). Für diesen Ansatz müssen jedoch ausreichend Daten vorhanden sein, um das KI-Modell zu trainieren, was bei einer Anwendung auf Fahrzeugebene zu prüfen wäre.

Eine weitere Herausforderung ist die Nutzung sogenannter Dynamic Changing Data. Im Normalfall werden die Daten, die es auszuwerten gilt, ständig durch neu erfasste Daten von Fahrzeugen ergänzt. Damit handelt es sich nicht um abgeschlossene, sich nicht mehr ändernde Daten (Static Data), wie sie beispielsweise oftmals im Gesundheitsbereich anfallen, sondern um einen sich ständig ändernden Datenbestand. Dies muss bei der Anonymisierung dieser Daten ebenfalls berücksichtigt werden. Unter Umständen kann sich die Stärke der Anonymisierung durch die ständig Erweiterung der Daten verändern.

Im Hinblick auf die strengen Richtlinien der GDPR dürfte im Falle der aggregationsbasierten Methoden ein weiteres Problem auftreten. Wie bereits beschrieben, gibt es für die aggregationsbasierten Methoden oftmals nur unzulängliche Bewertungen der Stärke der Anonymität, die bei diesen Methoden zudem immer von unbekannten Daten Dritter abhängt. Damit lässt sich vermutlich nur schwer beweisen, dass Daten, die mit diesen Methoden anonymisiert wurden, tatsächlich ausreichend geschützt sind. Gerade im Hinblick auf Informationen Dritter, die außerhalb der zu anonymisierenden Daten existieren und die die Stärke der Anonymisierung mit Hilfe von aggregationsbasierten Verfahren womöglich einschränken, verdeutlicht der Automobilbereich sehr klar, welche Probleme bei diesen Verfahren entstehen können. So kennen die Konzerne im Normalfall ihre Fahrer/Kunden und damit verbunden auch zahlreiche Informationen über diese wie Wohnort, Fahrzeugmodell, womöglich sogar den Arbeitgeber. Mit Hilfe dieser externen Informationen lassen sich Daten, die mit Hilfe aggregationsbasierter Methoden anonymisiert wurden, unter Umständen leicht re-identifizieren.

Zuletzt könnte es auch relevant werden, die Daten hinsichtlich des Herstellers selbst zu anonymisieren. Werden zum Beispiel herstellerübergreifend Daten zusammengeführt, um diese für verschiedene Anwendungs- und Forschungszwecke zu verwenden, kann es auch nötig sein, die Daten so zu anonymisieren, dass kein Rückschluss mehr auf den jeweiligen Hersteller gezogen werden kann. In diesem Fall geht es weniger darum, persönliche Daten einzelner Nutzer zu anonymisieren, sondern darum, firmeneigene Merkmale der Daten zu schützen. Es wäre beispielsweise sicherlich nicht im Sinne eines Herstellers, wenn sich im Rahmen einer Analyse herausstellt, dass bei Autos dieses Herstellers im Vergleich zu anderen Herstellern deutlich mehr oder früher Probleme mit dem Motor auftreten.

Wie in Abschnitt 3.2.3 beschrieben, beschäftigen sich einige Anonymisierungsmethoden mit Graph Data. Diese könnten im Zusammenhang mit Fahrzeugen interessant werden, wenn es um die Kommunikation von Fahrzeugen untereinander oder auch um die Kommunikation von Fahrzeugen mit Infrastruktur (beispielsweise Ampeln, Messstation oder Verkehrsleitsystemen) geht. Die erhobenen Daten dürften vermutlich als Graph Data erhoben und gespeichert werden. Diese Netzwerke können mit Hilfe der in Abschnitt 3.2.3 beschriebenen Ansätze anonymisiert werden. Auch die kurz in Abschnitt 3.2.1 angesprochenen Ansätze zur künstlichen Datengenerierung könnten für die Zukunft - nicht nur im Automobilbereich - eine immer größere Rolle spielen. Der Ansatz bietet eine hohe Anonymisierung bei gleichzeitig hoher Datennutzbarkeit. Und die große Menge an Daten, die für das Trainieren solcher neuronalen Netze benötigt werden, dürfte im Automobilbereich ohne weiteres die nächsten Jahre erhoben werden.

5 Implementierung ausgewählter Ansätze

Nachdem in den vorherigen Kapiteln ausführlich über die theoretischen Aspekte der Datenanonymisierung diskutiert und verschiedene Vor- und Nachteile abgewogen wurden, soll in diesem Abschnitt nun ein Ansatz beispielhaft für einen Anwendungsfall im Automobilbereich implementiert werden. Ursprünglich war vorgesehen, verschiedene Anwendungsfälle und verschiedene Ansätze aus verschiedenen Anwendungsgebieten und auch mit unterschiedlichen Methodiken zu implementieren. Die Beschaffung entsprechender Daten gestaltete sich jedoch nicht so einfach (siehe Abschnitt 5.1), weshalb am Ende nur ein Ansatz für einen Anwendungsfall umgesetzt wurde. Nichtsdestotrotz lassen sich auch an diesem Beispiel einige interessante Aspekte darstellen.

5.1 Datengenerierung

In diesem Abschnitt soll zunächst auf die Generierung der Daten eingegangen werden, die anonymisiert werden sollen. Echte Daten können zu diesem Zweck nicht genutzt werden, da die Verwendung dieser Daten laut Datenschutz-Behörden nicht erlaubt ist. Um solche tatsächlich erhobenen Daten zu nutzen, muss für alle Daten die explizite Einverständniserklärung der Nutzer zur Nutzung der Daten für diesen Zweck vorliegen. Dies ist bei den meisten Daten nicht der Fall. Ohne Einverständniserklärung dürfen die Daten nur vollständig anonymisiert verwendet werden, was jedoch genau Teil dieser Arbeit und auch das Ziel der Implementierung ist.

Folglich müssen die zu anonymisierenden Daten entweder simuliert werden oder mit explizitem Einverständnis der Personen, die die Daten erzeugen, extra für diesen Zweck erhoben werden. Beide Möglichkeiten wurden in dieser Arbeit geprüft und werden in den folgenden Abschnitten besprochen.

5.1.1 Fahrzeugdaten mit CarMaker

Die Firma IPG Automotive bietet mit der Software CarMaker [1] ein Simulationsprogram an, das eine Vielzahl an verschiedenen Fahrzeugdaten simulieren kann. So lassen sich dort verschiedene Testszenarien, Fahrtstrecken, Fahrmanöver und Fahrzeuge auswählen, die auf Basis der spezifizierten Parameter eine große Menge Daten erzeugen und damit die Erfassung von Daten in einem realen Auto simulieren. Im Rahmen dieser Arbeit wurde das Programm mit Hilfe einer Testlizenz genutzt, um Daten zu generieren, die im Anschluss anonymisiert werden sollen. Es zeigte sich jedoch schnell, dass die erfassten Daten größtenteils auch im späteren tatsächlichen Betrieb nicht anonymisiert werden müssen, da es sich hauptsächlich um sehr technische Daten handelt, die keine persönlichen Daten darstellen. Zudem ist die Erstellung solcher Simulationen sehr aufwendig und benötigt eine gewissen Einarbeitungszeit. Hinzu kommt, dass für die Anonymisierung eine gewisse Menge an Daten unterschiedlicher Autos benötigt wird, sodass die Generierung dieser Daten mit Hilfe der Software CarMaker grundsätzlich wahrscheinlich möglich wäre, in der für diese Arbeit zur Verfügung stehenden Zeit jedoch nicht umsetzbar ist. Folglich wurde die Implementierung eines Ansatzes zur Anonymisierung von Daten, die mit Hilfe der Software CarMaker erzeugt wurden, verworfen.

5.1.2 GPS-Daten mit OBD

Um Routen oder andere GPS-Daten zu generieren, gibt es verschiedene kommerzielle Tools. Um jedoch möglichst realistische Daten zu erzeugen, wurde im Rahmen dieser Arbeit ein sogenannter OBD GPS-Tracker genutzt. Diese GPS-Tracker können an die standardisierte OBD-Schnittstelle, über die alle Fahrzeuge, die ab 2004 produziert wurden, verfügen, angeschlossen werden. Über diese Schnittstelle werden unter anderem in Werkstätten die Fahrzeugdaten ausgelesen. Der GPS-Tracker beinhaltet eine SIM-Karte und zeichnet, sobald das Gerät angeschlossen ist, in regelmäßigen Abständen die aktuelle Position auf.

Für diese Arbeit wurden verschiedene Personen ausgewählt, die über einen bestimmten Zeitraum einen solchen GPS-Tracker in ihrem Auto angeschlossen hatten. Somit konnten über einen Zeitraum von 40 Tagen insgesamt 30.868 Datenpunkte von vier verschiedenen Personen bzw. Fahrzeugen aufgezeichnet werden.

Abbildung 5.1 zeigt die mit dem GPS-Tracker aufgezeichneten Routen für eine Woche in dem Erfassungszeitraum. Über eine Online-Schnittstelle lassen sich alle aufgezeichneten Werte und Routen darstellen und exportieren. Somit können die erfassten Daten im nächsten Schritt anonymisiert werden.

Die aufgezeichneten Daten werden im Folgenden genutzt und mit Hilfe eines ausgewählten Verfahrens anonymisiert. Zwar sind die Daten echt und wurden tatsächlich aufgezeichnet. Trotzdem weichen die Daten von den später tatsächlich verwendeten Daten ab. So ist auf

[1] https://ipg-automotive.com/de/produkte-loesungen/software/carmaker/

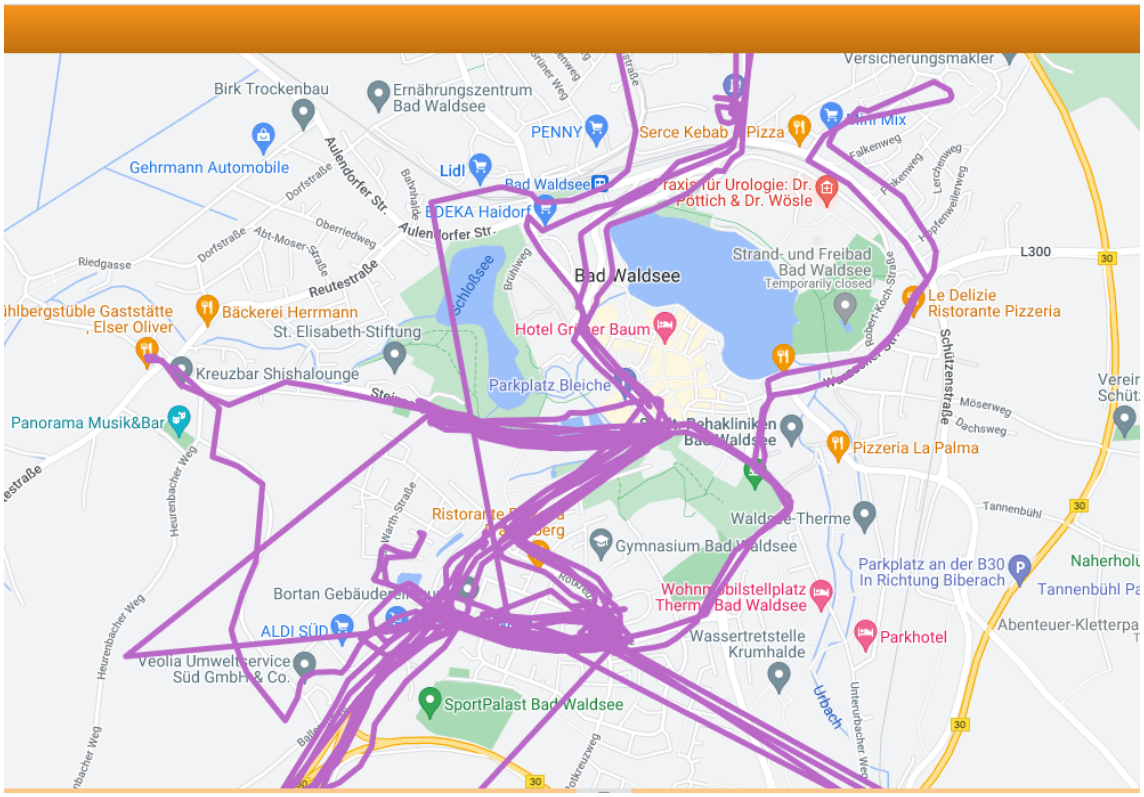

Abbildung 5.1: Visualisierung der mit Hilfe des OBD GPS-Trackers aufgezeichneten Routen in einer Woche.

Grund des Einsatzes von nur einem GPS-Trackers immer nur ein Datenpunkt gleichzeitig aufgezeichnet. Bei mehreren GPS-Trackern wären zu einem bestimmten Zeitpunkt immer auch mehrere Datenpunkte verfügbar. Zudem ist bei den erzeugten Daten auf Grund des Zeitstempels bereits eindeutig identifizierbar, von welcher Person die Daten stammen. Zwar ist nur wenigen Personen bekannt, wer zu welcher Zeit den GPS-Tracker benutzte. Wie jedoch in Kapitel 2 beschrieben wurde, dürfen anonymisierte Daten nicht mit Hilfe von externem Wissen re-identifizierbar sein. In diesem Fall wäre das für einige Personen jedoch möglich, was dazu führt, dass die Daten als nicht anonymisiert gelten.

5.2 Auswahl des Papers

Für die Auswahl eines Ansatzes zur Implementierung der oben geschilderten Daten galt es einige Dinge zu berücksichtigen. Zunächst soll ein Ansatz ausgewählt werden, der Routen, also Trajectories, anonymisiert. Zudem soll der vorgestellte Ansatz nachvollziehbar im Paper dargestellt sein und die verwendeten Algorithmen ausreichend verständlich erklärt sein. Einige Paper skizzieren nur grob die Idee, eine detaillierte Beschreibung des Ansatzes ist jedoch in einigen Arbeiten nicht zu finden. Zudem sollte das Paper möglichst entweder

das Konzept der Differential Privacy nutzen oder einen aggregationsbasierten Ansatz beschreiben. Wie in Abschnitt 3.2.3 erläutert, sind dies die meist genutzten Methodiken für diesen Datentyp. Auf diese soll in der Implementierung auch eingegangen werden.

Zuletzt wurde bei der Sichtung der verschiedenen Paper deutlich, dass viele Ansätze auf der Grundannahme beruhen, dass die aufgezeichneten und zu anonymisierenden Daten aus einem Triple bestehen: Latitude, Longitude und der Zeitstempel. Es wird davon ausgegangen, dass du einem bestimmten Zeitpunkt t eine gewisse Anzahl an Datenpunkten verschiedener Personen existieren. Diese Grundannahme ist auf Grund der Art und Weise, wie die Daten für die Implementierung erhoben wurden, nicht gegeben. Es gibt zu einem bestimmten Zeitpunkt jeweils nur genau einen Datenpunkt, da nur ein Messgerät genutzt wurde. Zwar ließen sich die Daten durch die Entfernung des Datums und nur der Beibehaltung der Zeit so manipulieren, dass zu einer bestimmten Zeit mehrere Datenpunkte vorliegen. Zum einen bedeutet das jedoch eine starke Veränderung der Ursprungsdaten. Zudem ist durch die Art und Weise, wie die Routen erzeugt werden (siehe Abschnitt 5.3.1) immer noch nicht sicher, dass zu jedem Zeitpunkt eine Mindestanzahl an Datenpunkten existiert.

Unter Berücksichtigung der genannten Punkte wurde für die Implementierung in dieser Arbeit der Ansatz von Rajesh et al. „Personalized trajectory anonymization through sensitive location points hiding“ aus dem Jahr 2019 ausgewählt [RAD19]. Die Grundidee des Ansatzes ist es, so genannte Haltingpoints in Routen zu identifizieren, an denen die jeweilige Person mindestens eine gewissen Zeit bleibt. Diese werden als wichtig und persönlich identifizierbar angesehen und deshalb anonymisiert. Dazu werden verschiedene Haltingpoints verschiedener Routen genutzt und mit Hilfe dieser ein MBR erstellt, durch das der tatsächliche Punkt ersetzt wird. Dadurch werden die für die Anonymität kritischen Punkte verallgemeinert und durch einen Bereich ersetzt, während die unkritischen Punkte der Route nach wie vor erhalten bleiben. Der Ansatz wurde für die Anonymisierung von Routenverläufen erstellt, die von Mobilgeräten aufgezeichnet werden. Es geht somit nicht direkt um Fahrzeugdaten, sondern um die Anonymisierung von Fußgängerdaten in Fußgängerzonen, Einkaufszentren oder sonstigen Gebieten.

5.3 Implementierung

Der im vorherigen Abschnitt beschriebene Ansatz wurde in Python implementiert. Da im Paper einige Punkte nicht eindeutig dargestellt sind oder ganz weggelassen wurden, wurden an diesen Stellen eigene Entscheidungen getroffen und umgesetzt. Eine exakte Umsetzung des Algorithmus wie der in der Arbeit beschrieben wurde, war nicht möglich.

Algorithmus 5.1 Datenaufbereitung

```
procedure CREATETRAJECTORIES(datapoints)
    for all point ∈ datapoints do
        if SAMEPOSITIONS(previousPoint,point) then
            Set new duration of previous point
            if duration > threshold then
                Finish route and save as trajectory
            end if
        else if POSITIONVALID(previousPoint,point) then
            if noRouteRecording then
                start new route
            end if
            append point to route
        end if
    end for
end procedure
```

5.3.1 Datenaufbereitung

Über die Online-Schnittstelle des OBD GPS-Tracker Herstellers wurden sämtliche Daten heruntergeladen und als CSV-Daten gespeichert. Die Daten enthalten neben der GPS-Position (Latitude und Longitude) auch einen Zeitstempel, eine aktuelle Geschwindigkeit sowie ein Validierungsstatus (boolean).

Um die Daten nun zu anonymisieren, wird zunächst definiert, wie im Folgenden der Begriff Trajectories (Routen) verwendet wird. In [RAD19] wird die Trajectory definiert als eine Liste von Tripeln, die jeweils aus der Latitude, der Longitude und dem Zeitstempel bestehen. Diese Definition ist auch in anderen Papern zu finden. Wann eine Route jedoch startet und endet, wenn Positionsdaten durchgängig aufgezeichnet wurden, wird nirgends definiert. Da die mit Hilfe des GPS-Trackers aufgezeichneten Daten praktisch eine lange Liste von Tripeln sind und lediglich als eine einzige Trajectory genutzt werden könnte, wird für diese Implementierung eine weitere Definition benötigt.

Um aus den aufgezeichneten Daten mehrere Trajectories zu generieren, die im Anschluss anonymisiert werden können, wird für diese Arbeit eine Trajectory bzw. Route folgendermaßen definiert: Eine Folge von Tripeln, für die gilt, dass die Dauer, für die sich die Position nicht ändert, eine bestimmte Zeit t nicht überschreiten darf. Für die Implementierung wurde

$$t_{routeThreshold} = 3600s$$

gewählt.

Des Weiteren wurde festgelegt, dass Routen Einträge mit mindestens n unterschiedlichen Positionsdaten benötigen. Für die Implementierung wurde

$$t_{minStops} = 10$$

gewählt. Damit wird ausgeschlossen, dass kurzzeitige Änderungen durch Messungenauigkeiten direkt als Route erkannt werden.

Da im Verlauf der Anonymisierung der Zeitstempel nicht mehr direkt benötigt wird, sondern nur noch die Dauer des Aufenthalts an einer bestimmten Position, wurden zudem alle Punkte, die dieselben Positionsdaten enthalten zusammengefasst und mit Hilfe des Zeitstempels die Aufenthaltsdauer an diesem Ort berechnet und gespeichert. Da wie bereits in Abschnitt 5.1.2 beschrieben bei diesem Szenario der Zeitstempel als eindeutiger Identifier gilt, muss dieser im Verlauf der Anonymisierung sowieso entfernt werden.

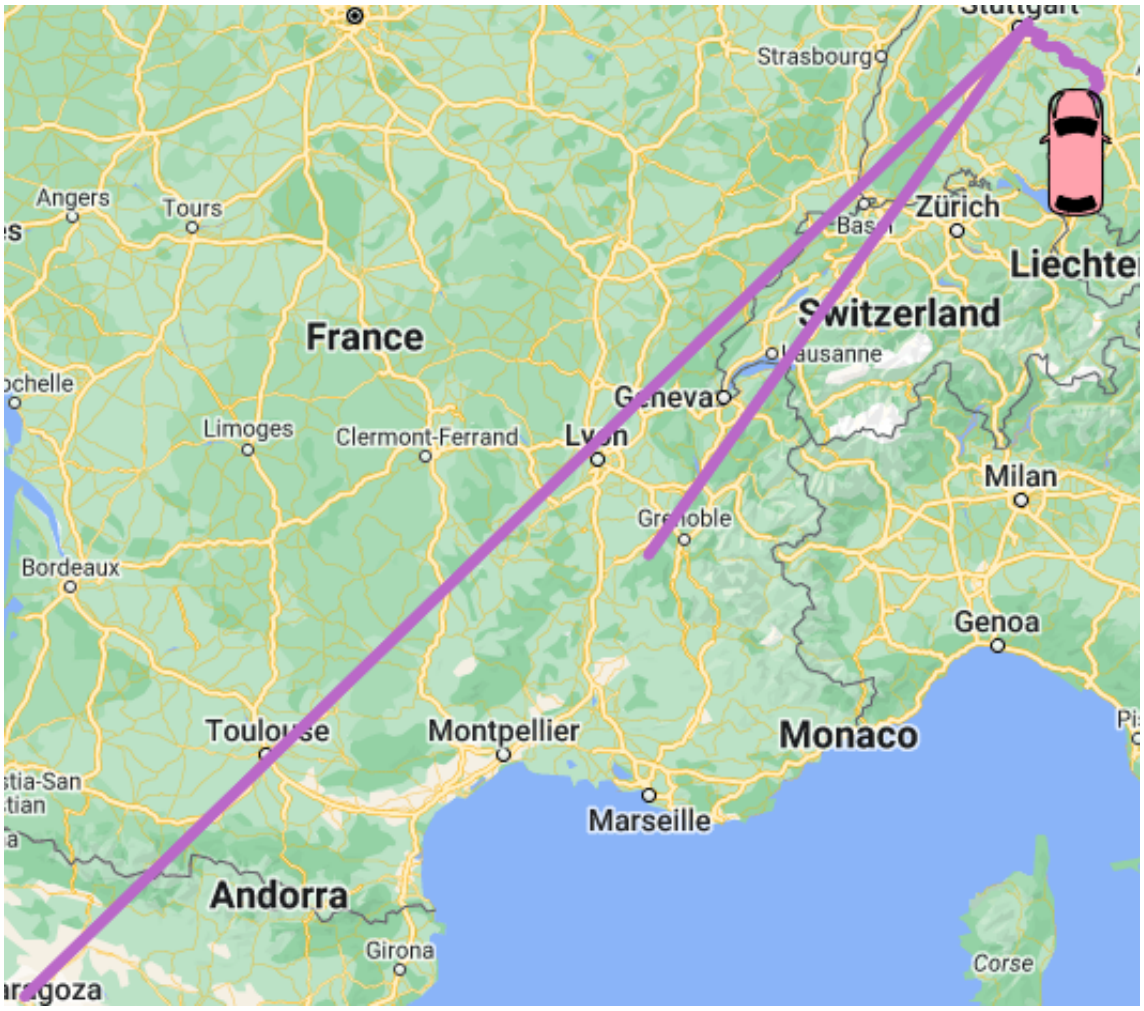

Abbildung 5.2: Ausreißer in den aufgezeichneten GPS-Daten

Algorithmus 5.2 Anonymisierung

```
procedure doAnonymization(trips)
    getHaltingpoints(trips)
    searchAndCalculateMBRs
    return anonymized trips
end procedure
```

Zuletzt müssen bei der Verarbeitung der Daten Ausreißer identifiziert und eliminiert werden. Abbildung 5.2 zeigt die aufgezeichneten Daten für einen bestimmen Zeitraum. Es ist klar zu sehen, dass die Messpunkte bei Zaragoza und Grenoble keineswegs Teile einer Route sind, sondern offensichtlich von fehlerhaften Messungen kommen. Um diese Ausreißer zu ignorieren, wird bei allen Punkten geprüft, ob der Abstand zum vorherigen Punkt einen bestimmten Schwellenwert t nicht überschreitet. Zur Messung der Distanz wurde der Euklidische Abstand genutzt. Für die Implementierung wurde

$$t_{maxDistance} = 0.5$$

gewählt. Algorithmus 5.1 zeigt den Ablauf zur Extrahierung der Routen.

Die Verarbeitung und Aufbereitung der Daten ergibt insgesamt 49 Routen. Dabei wurden 714 Datenpunkte als Ausreißer definiert und ignoriert. Bei insgesamt rund 30.000 Datenpunkten entspricht das 0,43%. Die 49 extrahierten Routen werden anschließend mit Hilfe des in [RAD19] beschriebenen Verfahrens anonymisiert.

5.3.2 Anonymisierung

Die Anonymisierung wurde vom Grundablauf wie im Paper skizziert ausgeführt. Jedoch wurde die Aufzeichnung der Preprocessing time und der Processing time weggelassen, da auf Grund der geringen Datenmenge ein Vergleich von Ausführungszeiten wenig Sinn ergibt.

Wie in Algorithmus 5.2 dargestellt, besteht die Anonymisierung im Wesentlichen aus zwei Schritten. Zunächst findet die Berechnung aller Haltingpoints statt. Anschließend wurden mit Hilfe dieser Haltingpoints die MBR berechnet.

Zur Berechnung der Haltingpoints wurden alle Datenpunkte aller Routen überprüft und die Aufenthaltsdauer betrachtet. Je nach Länge der Aufenthaltsdauer wurden die Punkte als sensitive oder non-sensitive Haltingpoints markiert. Anschließend wurden alle sensitiven Haltingpoints wie im Paper beschrieben nach Aufenthaltsdauer absteigend sortiert. Für die Implementierung wurde der Schwellenwert für die Einordnung als sensitiver Haltingpoint

$$t_{sensitive} = 60s$$

gewählt. Schwellenwert für die Einordnung als non-sensitiver Haltingpoint wurde mit

$$t_{nonSensitive} = 25s$$

Algorithmus 5.3 Berechnung Haltingpoints

```
procedure getHaltingpoints(trips)
    sensitiveHaltingPoints = []
    nonSensitiveHaltingPoints = []
    for all trip ∈ trips do
        for all point ∈ trip do
            if point.duration > sensitiveThreshold then
                point.markAsSensitive()
            else if point.duration > nonsensitiveThreshold then
                point.markAsNonSensitive()
            end if
        end for
    end for
    sensitiveHaltingPoints.sort()                    // by duration descending
    return sensitiveHaltingPoints, nonSensitiveHaltingPoints
end procedure
```

gewählt. Algorithmus 5.3 zeigt den Ablauf der Berechnung der Haltingpoints.

Mit diesen Schwellenwerten wurden in den 49 Routen insgesamt 159 sensitive Haltingpoints und 64 non-sensitive Haltingpoints gefunden. Die Start- und Endpunkte jeder Route sind jeweils sensitive Haltingpoints. Somit wurden noch rund 50 weitere sensitive Haltingpoints auf der zurückgelegten Route gefunden. Die in dem Paper vorgeschlagenen Schwellenwerte von 120s bzw. 150s wurden nach unten korrigiert, damit genügend Haltingpoints außerhalb der Start- und Endpunkte gefunden werden. In Abschnitt 5.3.3 wird auch nochmals auf die Bewertung und die Schwellenwerte der Haltingpoints eingegangen.

Um im Anschluss die MBRs zu bestimmen, werden für jeden sensitiven Haltingpoint die n nahesten Haltingpoints (sensitive oder non-sensitive) berechnet. Für n wurde in der finalen Version n=3 gewählt. Das bedeutet, dass die MBR jeweils den sensitiven Haltingpoint sowie drei weitere Punkte einschließen. Zur Berechnung der Distanz zwischen Punkten wurde nicht wie im Paper beschrieben die Haversine Distance genutzt, sondern wie bereits bei der Vorverarbeitung der Daten der Euklidische Abstand. Die Nutzung der Haversine Distance führte zu unbrauchbaren Ergebnissen, was jedoch vermutlich an einer fehlerhaften Implementierung lag. Da der Unterschied zum Euklidischen Abstand bei der kurzen Distanz jedoch irrelevant sein dürfte, wurde dieser für die Implementierung genutzt.

Im Paper ist nicht eindeutig beschrieben, ob in dem jeweiligen MBR lediglich ein sensitiver Punkt und ansonsten ausschließlich nicht-sensitive Punkte enthalten sein sollen, oder ob auch mehrere sensitive Punkte berücksichtigt werden können. Die Nutzung ausschließlich nicht-sensitiver Punkte führte jedoch zu sehr schlechten Ergebnissen, weshalb im Folgenden davon ausgegangen wird, dass für die Berechnung des MBR der sensitive Punkt und wie bereits beschrieben n weitere sensitive oder nicht-sensitive Punkte genutzt werden. Um die Privacy wie in Schritt 9.b) des Algorithmus („Increase NSP_n for more privacy and

Algorithmus 5.4 Berchenung Punkte für MBRs

```
procedure searchAndCalculateMBRs(sensPoints,nonSensPoints,n)
    allPoints = sensPoints + nonSensPoints
    for all sensPoint ∈ sensPoints do
        distanceList = []
        for all point ∈ allPoints do
            if point <> sensPoint then
                distanceList.append(point, euclideanDistance(point,sensPoint))
            end if
        end for
        distanceList.sort()                                  // by distance ascending
        pointList = [sensPoint]
        for all i → n do
            pointList.append(distanceList[i])
        end for
        sensPoint.setMBR(new MBR(pointList))
    end for
end procedure
```

enlarge Z_A“, [RAD19]) zu erhöhen, wird damit nicht die Anzahl an non-sensitiven, sondern die gesamte Anzahl an berücksichtigten Punkten erhöht. In Algorithmus 5.4 wird die Berechnung der Punkte für das jeweilige MBR der sensitiven Punkte beschrieben. Das MBR selbst wird als eigenes Objekt gespeichert. Mit Hilfe der übergebenen Punkte wird das MBR berechnet und kann im Anschluss die ursprünglichen Koordinaten des sensitiven Punkts ersetzen. Die Implementierung ermöglicht im Anschluss sowohl die Ausgabe des kompletten MBR, als auch die Ausgabe des Mittelpunkts des MBR, um einen einzelnen Punkt für weitere Auswertungen nutzen zu können. In Algorithmus 5.5 ist nur die Berechnung des MBR beschrieben. Die Rückgabe der Koordinaten ist trivial und wird nicht näher beschrieben.

Die Berechnung des MBR wurde in der Implementierung vereinfacht. Ein typischer Ansatz zur Berechnung des MBR ist der „Rotating Calipers Algorithm“. Es wird zuerst die konvexe Hülle der Punkte berechnet und anschließend ein minimales Rechteck mit jeweils einer Strecke zwischen zwei Punkten als Kante des Rechtecks berechnet [Ebe]. Da die Berechnung jedoch relativ komplex ist und es für den Anwendungsfall in diesem Paper keine Rolle spielt, ob das Rechteck tatsächlich das kleinstmögliche ist, oder minimal größer, wurde für die Implementierung ein vereinfachter Ansatz gewählt. Es wird jeweils der kleinste bzw. größte Wert aller zu berücksichtigenden Punkte der jeweiligen Achse berechnet. Die Kanten des dadurch entstehenden MBR sind jeweils parallel zur X- und Y-Achse.

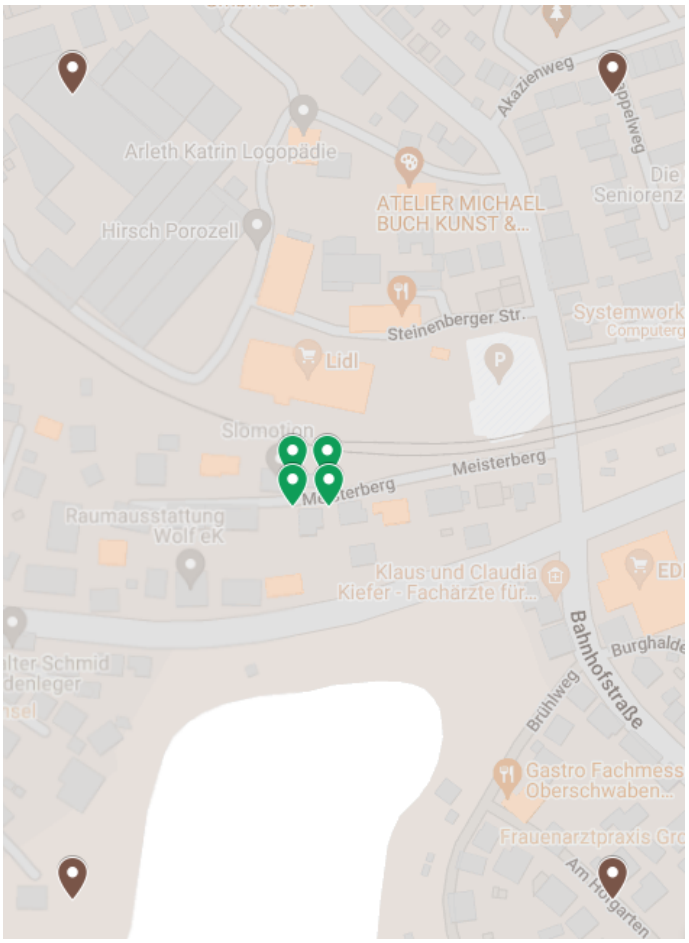

Abbildung 5.3: Vergleich des eigentlichen MBR mit dem Rechteck mit Mindestgröße

Je größer ein MBR ist, desto höher ist die Anonymität, da der mögliche Bereich, in dem die Punkte liegen, größer wird. Da sehr kleine MBR kaum Anonymität bieten, wurde für die Implementierung abweichend vom Paper eine Mindestgröße des MBR festgelegt. Für die Implementierung wurde

$$t_{minMBRArea} = 0.004$$

gewählt. Abbildung 5.3 zeigt die Problematik bei zu kleinem MBR. Das grüne Rechteck ist in einem Beispiel das ursprünglich berechnete Rechteck. Innerhalb dieses Rechtecks befinden sich 4 Haltingpoints, die alle sehr nahe beieinander sind (beispielsweise unterschiedliche Parkplätze an diesem Ort). Das daraus entstehende MBR verbirgt zwar den exakten Standort, lässt eine Lokalisierung jedoch immer noch sehr genau zu. Das braune Rechteck wurde unter Berücksichtigung der Mindestgröße generiert. Hier ist eine gewisse Anonymität gewährleistet.

Zuletzt werden alle sensitiven Punkte einer Route durch die MBR ersetzt. Im Paper wird nicht genau beschrieben, wie die jeweiligen Einträge genau ersetzt werden. In dieser Implementierung wurden die Koordinaten der sensitiven Punkte durch die Koordinaten

Algorithmus 5.5 Berechnung MBRs

```
class MBR
    points: A,B,C,D
    procedure init(basepoints)
        list_lat,list_lon = []
        for all p ∈ basepoints do
            list_lat.append(p.lat)
            list_lon.append(p.lon)
        end for
        A = (MAX(list_lat),MIN(list_lon))
        B = (MAX(list_lat),MAX(list_lon))
        C = (MIN(list_lat),MIN(list_lon))
        D = (MIN(list_lat),MAX(list_lon))

        if width < MIN_WIDTH then
            Enlarge width
        end if
        if heigth < MIN_HEIGHT then
            Enlarge height
        end if
    end procedure
end class
```

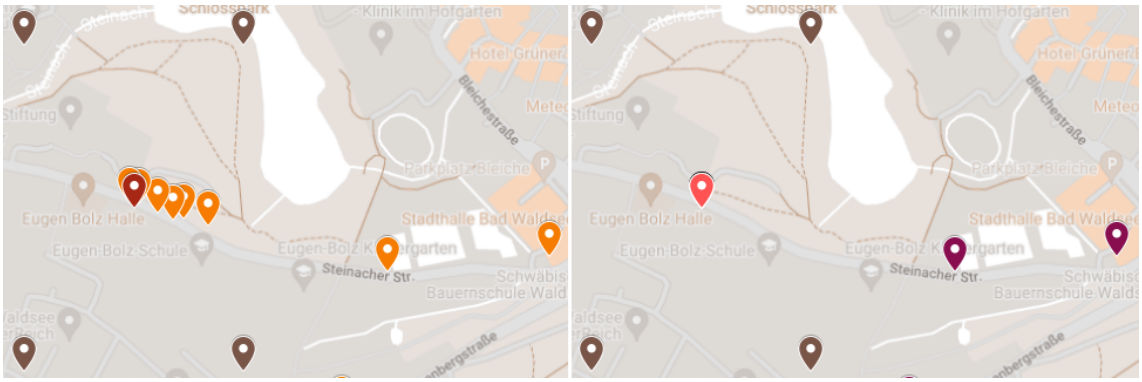

Abbildung 5.4: Links: Nur Anonymisierung des sensitiven Punkts. Rechts: Anonymisierung aller Punkte, die im MBR liegen

des Mittelpunkts des MBR ersetzt. Hierbei zeigte sich ein weiteres Problem bei der Umsetzung des im Paper beschriebenen Ansatzes. So werden zwar die sensitiven Punkte - und damit typischerweise die Start- und Endpunkte der Route - durch das MBR bzw. dessen Mittelpunkt ersetzt, die letzten Punkte bis zur Route werden jedoch nicht verfremdet. So zeigt die linke Abbildung in Abbildung 5.4 deutlich, dass zwar der Endpunkt anonymisiert ist (roter Punkt als Mittelpunkt des braunen Rechtecks), alle Punkte bis zu diesem Punkt jedoch nicht. Damit ist trotzdem deutlich zu sehen, wo die Route endet.

Algorithmus 5.6 Ersetzen aller Punkte in MBRs

```
procedure REPLACEALLMBRs(trip)
    mbrList = []
    for all stop ∈ trip do
        if stop.isSensitivePoint then
            mbrList.append(stop.mbr)
        end if
    end for
    for all stop ∈ trip do
        for all mbr ∈ mbrList do
            if POINTINMBR(stop,mbr) then
                stop.mbr = mbr                          // Add MBR to point
            end if
        end for
    end for
end procedure
```

Um diese Problematik zu umgehen, wurde in dieser Implementierung zuletzt für alle Punkte, die vor/nach den sensitiven Punkten liegen, geprüft, ob diese im MBR des sensitiven Punkts liegen. Ist dies der Fall, wurden die Punkte ebenfalls durch das MBR ersetzt. Algorithmus 5.6 zeigt die Ersetzung der Punkte, die in den MBR liegen. Die rechte Abbildung in Abbildung 5.4 zeigt die Route nach der Ersetzung der angrenzenden Punkte. Alle Punkte, die ursprünglich innerhalb des MBR lagen, liegen nun im Zentrum des MBR. Damit lässt sich nicht mehr genau sagen, wo im MBR die Route endet.

5.3.3 Ergebnis und Bewertung

Darstellung

Mit Hilfe der oben beschriebenen Punkte konnten die aufgezeichneten Daten verarbeitet, Routen extrahiert und anonymisiert werden. Auf eine detaillierte Berechnung der Datennutzbarkeit, des Information Loss und der Anonymität wird hier verzichtet, da es zunächst um die grundsätzliche Umsetzbarkeit des Ansatzes mit Fahrzeugdaten ging. Die Datennutzbarkeit, das Anonymitätslevel und weitere Aspekte und Herausforderungen bei der Implementierung sollen jedoch im Folgenden näher beleuchtet und diskutiert werden.

Zur Veranschaulichung der Ergebnisse (die Originalrouten sowie die anonymisierten Routen) wurden die Daten als CSV-Datei exportiert. Google bietet mit Google My Maps [2] die Möglichkeit, eigene Karten zu erstellen. Unter anderem ist es möglich, CSV-Dateien mit GPS-Daten und weiteren Informationen zu importieren. Mit Hilfe dieser Karten war es

[2] https://www.google.com/intl/de_de/maps/about/mymaps/

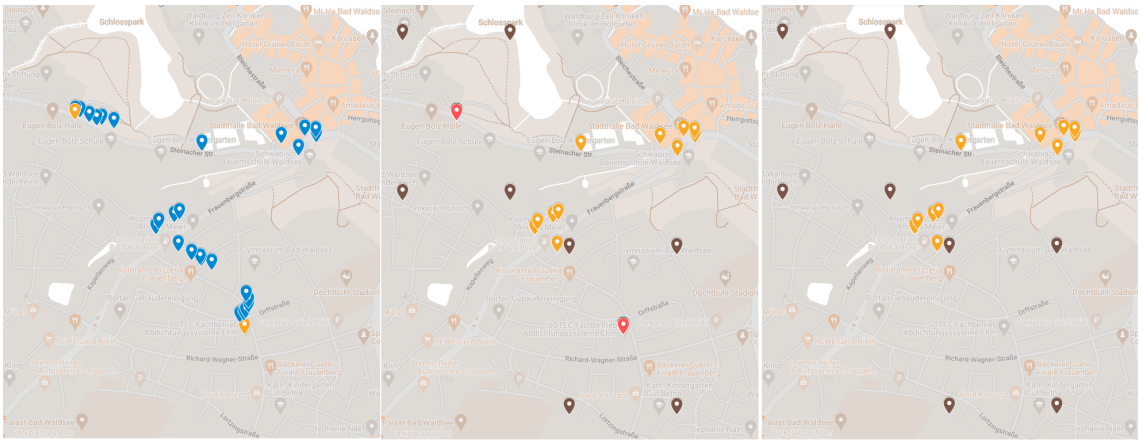

Abbildung 5.5: Links: Original Route. Mitte: Anonymisierte Route mit Zentrum des MBR. Rechts: Anonymisierte Route ohne Zentrum des MBR

möglich, verschiedene Routen und Zwischenschritte zu visualisieren. Abbildung 5.5 zeigt zunächst anhand eines Beispiels eine Originalroute, wie sie nach der Datenaufbereitung gespeichert ist. In der Mitte ist die final anonymisierte Route dargestellt. Es ist klar zu erkennen, dass die Start- und Endpunkte der Route sowie deren umgebende Punkte nicht mehr eindeutig identifizierbar sind. Da in diesem Fall das Zentrum des MBR jeweils nahe am Start- bzw. Endpunkt der Route liegen, zeigt die rechte Abbildung lediglich die MBR ohne Zentrum. Hierdurch wird deutlich, dass die Start- und Endpunkte tatsächlich nicht mehr genau identifizierbar sind.

Anonymität

Betrachtet man die Implementierung im Hinblick auf die Anonymität der verarbeiteten Routen, bietet sich ein differenziertes Bild. Zunächst ist festzustellen, dass eine gewisse Anonymität, wie im vorherigen Abschnitt beschrieben, erreicht wurde. Start- und Endpunkte sowie längere Aufenthalte im Verlauf der Route werden durch eine Fläche ersetzt, in der die entsprechenden Orte überall liegen können. Sorgt man dafür, dass die MBR groß genug sind, bietet diese Fläche eine gewisse Anonymität. Die Wirksamkeit der MBR hängt jedoch maßgeblich von den vorher berechneten Haltingpoints ab. Liegt beispielsweise der sensitive Punkt der Route, der durch das MBR ersetzt werden soll, auf der Kante des Rechtecks oder womöglich sogar auf der Ecke, da alle weiteren Punkte, die zur Berechnung des MBR genutzt werden, in der gleichen Richtung vom sensitiven Punkt entfernt liegen, können unter Umständen relativ einfach Rückschlüsse auf die tatsächliche Lage des sensitiven Punkts gezogen werden. Auch hat die Lage des MBR im Vergleich zur Route maßgeblich Auswirkungen auf die Stärke der Anonymität. Führt die Route komplett durch ein sehr großes MBR, so können nahezu keine Informationen über die ursprüngliche Route gewonnen werden. Umgekehrt können viele Informationen zur ursprünglichen Route gewonnen werden, wenn diese gar nicht durch ein MBR führt.

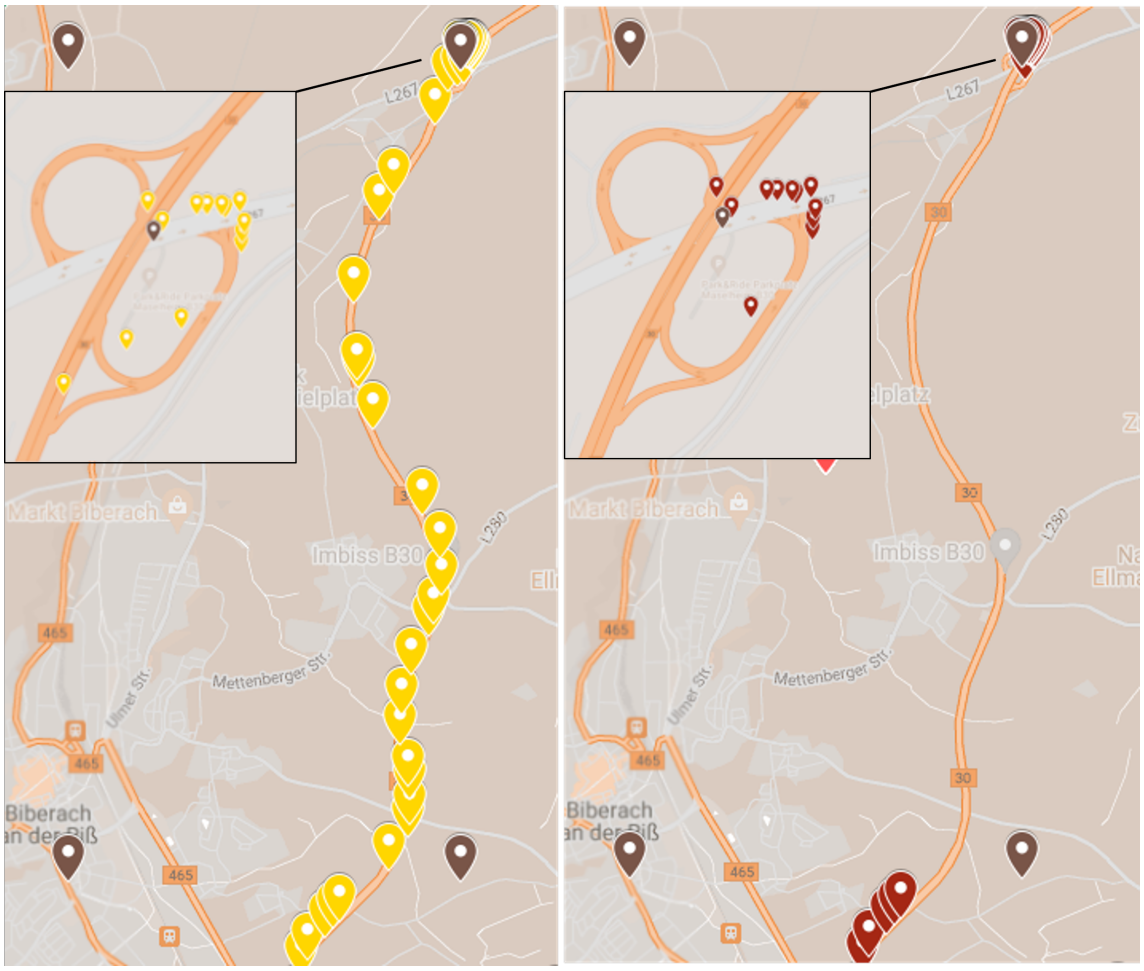

Abbildung 5.6: Links: Original Route. Rechts: Anonymisierte Route mit MBR

Abbildung 5.6 zeigt die beiden geschilderten Beispiele in einem einzigen Fall. Die Originalroute im Bild links führt durch das später berechnete MBR. Am nordwestlichen Rand verlässt die Route jedoch das MBR. Nur den Endpunkt und ein weiterer Punkt liegen wieder im MBR. Der Endpunkt ist gleichzeitig der nordwestlichste Punkt der nahegelegenen sensitiven und nicht-sensitiven Punkte. Im rechten Bild ist erkennbar, dass die Route im MBR durch das MBR ersetzt wird. Die Punkte an der oberen rechten Ecke lassen jedoch sehr deutlich erkennen, dass der Endpunkt vermutlich auf dem Parkplatz am Rand des MBR war.

Wie bereits im Implementierungsabschnitt beschrieben, birgt die Gruppierung vieler sensitiver und non-sensitiver Punkte die Gefahr, dass MBR zu klein werden und keine ausreichende Anonymität bieten. Für diese Implementierung wurde die Problematik damit gelöst, dass das MBR eine Mindestgröße haben muss, die definiert werden kann. Ohne diese zusätzliche Einschränkung wäre die Stärke der Anonymität in einigen Fällen viel zu gering. Es müsste jedoch geprüft werden, ob dieses Problem auch anders gelöst werden könnte. Beispielsweise wäre es möglich, sensitive Punkte, die in einer gewissen Nähe

zueinander liegen, zusammenzufassen. Wenn eine Route zum Beispiel auf einem Parkplatz endet, wird voraussichtlich jedes Mal eine andere Position (ein anderer Parkplatz) gemessen. Semantisch wäre der Parkplatz an sich jedoch die korrekte Endposition. Die Umsetzung ist jedoch komplex, da somit nach einer ersten Iteration, in der alle Haltingpoints berechnet werden, eine zweite folgen müsste, um nun alle zusammengefassten Haltingpoints in den Routen zu ersetzen.

Eine große Problematik, die dieser Ansatz gar nicht beachtet, ist die Eindeutigkeit längerer Routen, die nur von einzelnen Individuen genutzt werden. Gibt es nur eine einzige Route zu einem Ziel, so kann diese trotz Anonymisierung der Start- und Endpunkte, eventuell eindeutig einer Person zugeordnet werden. Zudem dürfte auf Grund fehlender sensitiver und non-sensitiver Punkte in der Nähe des Endpunktes ein sehr großes MBR entstehen, das fast die komplette Route verdeckt und die Daten damit nicht nutzbar macht. Im Falle dieser Implementierung kommt dies nicht vor, da auf Grund von Messungenauigkeiten in der Nähe des Zielortes ausreichend sensitive Punkte gefunden wurden. Diese Route müsste jedoch im Sinne der Anonymisierung unterdrückt werden, da sie auf Grund fehlender ähnlicher Daten nicht anonymisierbar ist.

Die Stärke der Anonymität lässt sich wie bei den meisten aggregationsbasierten Verfahren nur schwer messen. In diesem Fall hängt die Stärke maßgeblich vom Wissen der beteiligen und in den Daten erfassten Personen ab. Durch zusätzliches Wissen über beispielsweise den Arbeitsplatz einer Person, deren Daten anonymisiert wurden, lässt sich relativ einfach eine Verbindung zwischen Person und Route herstellen. Dieses mögliche zusätzliche Wissen lässt eine Bewertung der Anonymität nur schwer zu.

Datennutzbarkeit

Grundsätzlich sind die Daten nach der Anonymisierung weiterhin sehr gut nutzbar. Dadurch, dass lediglich die Start- und Endpunkte sowie Punkte auf der Route, an denen eine längere Pause eingelegt wurden, verfremdet werden, bleibt ein beträchtlicher Anteil der Route im Originalzustand. Damit ist die Datennutzbarkeit für diese Abschnitt sehr hoch, respektive der Information Loss 0. In den Bereich des Start- und Endpunktes ist die Datennutzbarkeit eher gering. Abhängig davon, wofür die Daten anschließend genutzt werden sollen, kann das eine mehr oder weniger große Rolle spielen. Geht es darum, grundsätzliche Verkehrsströme zu analysieren und zu prüfen, welche Wege häufig befahren werden, dürfte die Nutzbarkeit der Daten sehr hoch sein. Geht es darum, konkrete Verkehrswege beim Start oder Ende der Route zu identifizieren, wäre dieser Anonymisierungsansatz vermutlich nicht sehr nützlich.

Zuletzt ist die Nutzbarkeit der Daten stark von den genutzten Parametern abhängig. So wirkt sich die Mindestgröße des MBR massiv auf die Anzahl verallgemeinerte Punkte in der Route aus. Auch haben die Schwellenwerte für die Einordnung als sensitiver oder non-sensitiver Punkt große Auswirkungen auf die Nutzbarkeit, da eine höhere oder geringere Anzahl an Haltingpoints direkte Auswirkungen auf die Generierung der MBR hat. Um

konkrete Aussagen über die Nutzbarkeit machen zu können, müsste ein Anwendungsfall entwickelt werden und die Daten einmal ohne und einmal mit Anonymisierung für diesen Anwendungsfall genutzt werden.

Weitere Erkenntnisse und Probleme

Grundsätzlich bleibt festzuhalten, dass die Ungenauigkeit der GPS-Daten große Auswirkungen auf das Anonymisierungsverfahren haben. Durch kleine Unterschiede bei der Erfassung der sensitiven Punkte können sich große Änderungen der MBR und damit der Stärke der Anonymität ergeben. Im Falle dieser Implementierung könnte die Ungenauigkeit der GPS-Daten ggfs. vom OBD kommen. Es bleibt zu prüfen, ob die GPS-Daten, die später in den Autos erfasst werden, genauer sind. Auch sind bei der Nutzung dieses OBD die Messintervalle nicht eindeutig. Die Intervalle variieren teilweise stark und es ist nicht klar erkennbar, wann GPS-Positionen gespeichert werden. Bei hochwertigeren Messgeräten sollten diese Probleme nicht mehr auftreten. Dennoch sollte bei einer späteren Implementierung zumindest geprüft werden, welche Zeitintervalle und wie genau die Daten erfasst werden.

Da leichte Ungenauigkeiten in den GPS-Daten auch in hochwertigeren Messgeräten nicht ausschließbar sind, ist zu prüfen, ob eine Korrektur der GPS-Daten nötig ist.

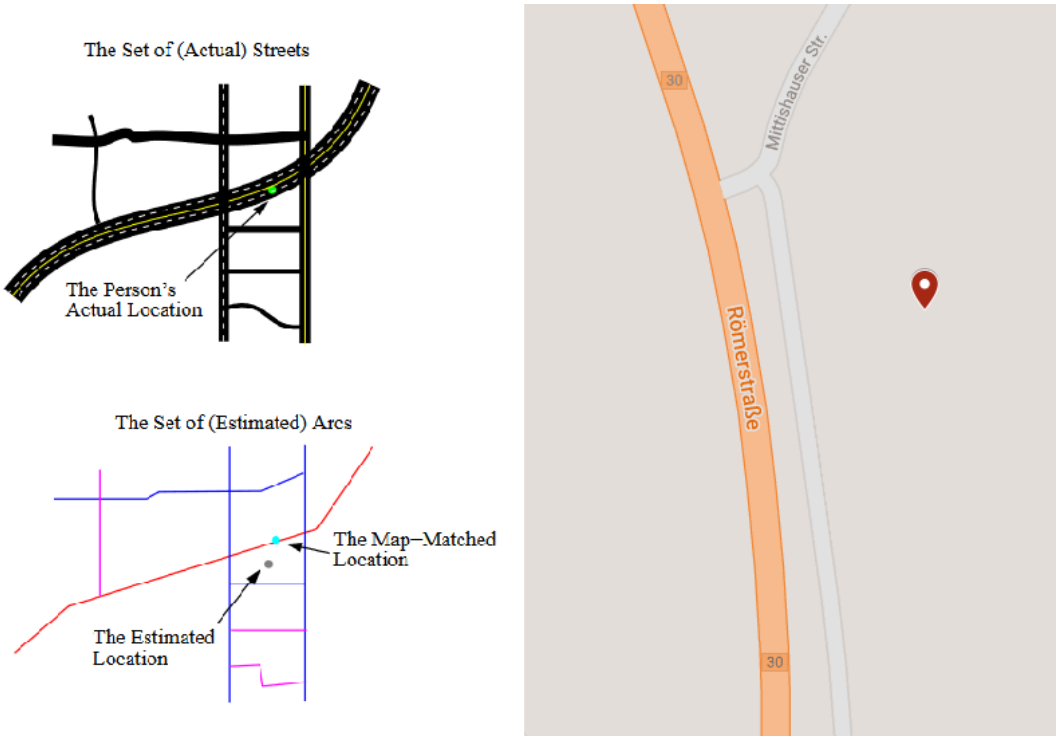

Abbildung 5.7: Links: Map-Matching um gemessene Positionen an Verkehrswege zu binden. Abbildung aus [BK96]. Rechts: Gemessener Punkt mit dem OBD

Abbildung 5.7 zeigt links die Problematik, gemessene Punkte an vorhandene Verkehrswege zu binden. Dieses Problem, dass gemessene Daten nicht auf Verkehrswegen liegen, bzw. das Projizieren der Daten auf vorhandene Verkehrswege wird „map matching problem“ genannt und in [BK96] näher behandelt. Links ist zu sehen, dass dieses Problem auch bei den Messungen des OBD auftritt. Ob eine solche Anpassung der Messpunkte an Verkehrswege tatsächlich nötig ist, hängt ebenfalls vom Anwendungsfall ab.

Zuletzt bleibt fraglich, ob das Konzept der Haltingpoints für Fahrzeugrouten Sinn ergibt. Start- und Endpunkte zu anonymisieren ist nachvollziehbar und trägt sicherlich zur Anonymisierung von Routen bei. Die berechneten non-sensitiven Haltingpoints ergeben im Zusammenhang mit Fahrzeugdaten jedoch wenig Sinn. Die in der Implementierung erfassten non-sensitiven Haltingpoints sind vermutlich Wartezeiten an Ampeln oder anderen Verkehrshindernissen, die mit den erfassten sensitiven Haltingpoints nichts zu tun haben. Sie erhöhen lediglich die Zahl der Punkte, die zur Berechnung der MBR genutzt werden können. Da die non-sensitiven Haltingpoints jedoch willkürlich und offensichtlich abhängig von der Verkehrslage entstehen, muss hinterfragt werden, ob eine Anonymisierung mit ausschließlich sensitiven Haltingpoints nicht sinnvoller wäre.

Zusammenfassend kann gesagt werden, dass die Implementierung eines Ansatzes für einen anderen Anwendungsfall (Anonymisierung von Fußgängerdaten in Fußgängerzonen oder Einkaufszentren) für die Automobildaten grundsätzlich möglich war und eine Umsetzung zu nutzbaren Ergebnissen geführt hat. Jedoch zeigten sich Probleme durch fehlende Details bei der Erklärung des Algorithmus im Paper. Außerdem mussten einige Anpassungen bei der Implementierung vorgenommen werden, damit die Anonymisierung für Fahrzeugdaten Sinn ergibt. Dies dürfte auch für andere Ansätze gelten, die nicht direkt für Fahrzeugdaten entwickelt wurden. Eine Anpassung der Konzepte wird in nahezu allen Fällen notwendig sein.

6 Ergebnisse

In dieser Arbeit wurde mit Hilfe einer strukturierten Literaturrecherche ein Überblick über die zurzeit verbreitetsten Methoden und Ansätze zur Anonymisierung verschiedener Daten gegeben. Die geringe Anzahl an Literatur zur Anonymisierung speziell im Automobilbereich machte eine umfassende Literaturrecherche notwendig. Dabei zeigt sich, dass aggregationsbasierte Methoden (typischerweise k-anonymity Ansätze, ergänzt durch l-diversity und t-closeness) mit rund zwei Dritteln aller Ansätze klar führend sind. Die flexiblen Einsatzmöglichkeiten dieser Methode sowie die Anpassung der Methode auf andere Datentypen wie Graph Data, Positionsdaten, Streaming Data oder Transactional Data erklären die große Verbreitung dieses Ansatzes. Ein großer Anteil mit rund 20% machen zudem pertubationsbasierte Methoden aus. Insbesondere Ansätze, die Differential Privacy nutzen, sind vermehrt in der Literatur zu finden. Zwar ist Differential Privacy in der Theorie ein enorm mächtiger Ansatz mit sehr vielen Vorteilen, in der Praxis scheint jedoch die Umsetzbarkeit vor allem hinsichtlich der Datennutzbarkeit bei ausreichend hoher Anonymität nur beschränkt möglich zu sein.

Die meisten Ansätze sind zudem keinem konkreten Anwendungsbereich zuordenbar. Dies zeigt, dass nach wie vor an Verbesserungen der mathematischen zu Grunde liegenden Verfahren oder im Falle der aggregationsbasierten Methoden auch an Clustering-Algorithmen geforscht wird und damit die existierenden Verfahren ständig verbessert werden. Mit Sozialen Netzwerken und dem Gesundheitsbereich stammt ein weiterer großer Anteil aus naheliegenden Anwendungsgebieten. Während sich die Ansätze aus den Gesundheitsbereich hauptsächlich mit klassischen Tabellendaten beschäftigen, finden sich in den Ansätzen aus dem Bereich Soziale Netzwerke oftmals Verfahren zur Anonymisierung von Graph Daten, sowie teilweise auch Positionsdaten. Diese zwei Datentyp sind nach tabellarischen Daten, mit denen sich fast 50% aller Methoden beschäftigen, am relevantesten. Vor allem Ansätze zur Anonymisierung von Positionsdaten - nochmals unterscheidbar in Routen (Trajectories) und Orte (POI) - dürften auch für den Automobilbereich sehr interessant sein.

Interessante Erkenntnisse ergaben sich auch hinsichtlich des Vergleichs verschiedener Ansätze, vor allem neuer Ansätze mit bisherigen. Es gibt offensichtlich keinen Standard zum Vergleich von Anonymisierungsansätzen, weder für die Stärke der Anonymisierung noch für die Datennutzbarkeit. Die Datennutzbarkeit wird zwar in den meisten Arbeiten mit Hilfe des Information Loss berechnet, jedoch gibt es eine Vielzahl von Definitionen und Berechnungen des Information Loss, was einen Vergleich über mehrere Paper hinweg sehr

schwierig macht. Die Stärke der Anonymität wird hingegen oftmals gar nicht verglichen. Wird sie verglichen, dann mit speziell definierten Metriken, die sich von Paper zu Paper unterscheiden.

Für den Automobilbereich lassen sich einige Ansätze und Ideen sicherlich leicht übertragen. Eine Anonymisierung von Standortdaten wird zwangsläufig eine große Rolle spielen. Doch auch die Anonymisierung anderer Datentypen wie Streaming und Transactional Data können für den Automobilbereich interessant sein. Obwohl aggregationsbasierte Methoden klar in der Überzahl sind, könnte eine Umsetzung dieser Ansätze in einigen Anwendungsfällen jedoch schwierig sein, da ein formaler Nachweis, dass die Daten ausreichend stark anonymisiert sind, bei diesen Ansätzen schwierig ist. Differential Privacy Ansätze könnten bei Automobildaten eine größere Rolle spielen, da neben der klaren Definition der Stärke der Anonymität auch ausreichend Daten erhoben werden könnten um die Nutzbarkeit der Daten trotzdem ausreichend zu erhalten.

Zuletzt zeigte sich in der beispielhaften Implementierung eines Ansatzes, dass im konkreten Anwendungsfall viele unterschiedliche Herausforderungen warten, bevor mit der Anonymisierung überhaupt begonnen werden kann. Zudem stellt sich die Frage, ob Routen- oder Positionsdaten in manchen Fällen überhaupt ausreichend anonymisiert werden können. Vermutlich lässt sich eine Anonymisierung mit ausreichender Datennutzbarkeit nur in größeren Städten umsetzen. In ländlichen Gebieten dürfte es zumindest für Positionsdaten oft schwierig werden, eine ausreichende Anonymität zu gewährleisten.

7 Ausblick

Ein nächster Schritt zur Anonymisierung von Daten im Automobilbereich könnte die Betrachtung konkreter Anwendungsfälle, wie in Abschnitt 4.1 beschrieben, sein. Hierbei können verschiedene Ansätze beispielhaft implementiert und ausgewertet werden. Nötig dafür wäre zudem eine Erarbeitung eines Standards zum Vergleichen von Anonymsierungsverfahren. Dieser Schritt wäre vermutlich nicht nur für den Automobilbereich, sondern allgemein für die Anwendung in verschiedenen Gebieten sinnvoll. Damit ließen sich zukünftig verschiedene Ansätze einfacher, schneller und transparenter vergleichen und damit auch leichter Entscheidungen über die Nutzung einzelner Verfahren treffen. Zuletzt wäre es sicherlich auch von Vorteil, eine technische Definition von Anonymisierung zu erarbeiten. Die aktuellen Definitionen beziehen sich lediglich auf allgemeine Beschreibungen und teilweise unterschiedlich interpretierbare Begriffe. Nicht zuletzt deshalb gibt es zum einen immer wieder Probleme, dass Anonymisierung falsch verstanden wird bzw. falsche Verfahren als Anonymisierungsverfahren betrachtet werden. Zum anderen macht eine fehlende technische Definition auch die Nutzbarkeit von Verfahren bzw. Daten im Hinblick auf die Vorgaben von Datenschutzbehörden sehr schwierig.

Die immer größeren Datenmengen, die im Automobilbereich anfallen, kombiniert mit dem Willen und der Notwendigkeit, diese Daten auch auszuwerten, dürfte in den nächsten Jahren für enorme Fortschritte im Bereich Anonymisierung von Automobildaten sorgen. Denn letztlich ist es sowohl im Sinne der Automobilhersteller als auch der Fahrzeuginhaber, dass starke Anonymisierungsverfahren für diesen Bereich entwickelt und eingesetzt werden.

Literaturverzeichnis

[] „A clustering-based anonymization approach for privacy-preserving in the healthcare cloud“. In: (). DOI: 10.1002/cpe.6487. URL: https://onlinelibrary.wiley.com/doi/10.1002/cpe.6487 (besucht am 07. 11. 2022) (zitiert auf S. 32).

[11] *Information technology — Security techniques — Privacy framework*. Standard. Geneva, CH: International Organization for Standardization, Dez. 2011 (zitiert auf S. 15, 16).

[AAS+18] A. Alnemari, S. Arodi, V. Sosa, S. Pandey, C. Romanowski, R. Raj, S. Mishra. *Protecting infrastructure data via enhanced access control, blockchain and differential privacy*. Bd. 542. IFIP Advances in Information and Communication Technology. Pages: 125. 2018. 113 S. DOI: 10.1007/978-3-030-04537-1_7. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059017001&doi=10.1007%2f978-3-030-04537-1_7&partnerID=40&md5=56787da83b4623bcd4f251475367b922 (zitiert auf S. 25, 28).

[Alj21] R. Aljably. *Privacy Preserving Data Sharing in Online Social Networks*. Bd. 1415. Communications in Computer and Information Science. Pages: 152. 2021. 142 S. DOI: 10.1007/978-981-16-3150-4_13. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85111430101&doi=10.1007%2f978-981-16-3150-4_13&partnerID=40&md5=6c96dc1784ab8454c6709882286bc00a (zitiert auf S. 28, 32).

[ALWJ18] S. An, Y. Li, T. Wang, Y. Jin. „Contact Graph Based Anonymization for Geosocial Network Datasets“. In: Proceedings - 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC 2018. 2018, S. 132–137. DOI: 10.1109/BESC.2018.8697836. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065251410&doi=10.1109%2fBESC.2018.8697836&partnerID=40&md5=b991397089e11268a23554ca54810a8f (zitiert auf S. 30).

[AM22] A. Abbasi, B. Mohammadi. „A clustering-based anonymization approach for privacy-preserving in the healthcare cloud“. In: *Concurrency and Computation: Practice and Experience* 34.1 (2022). DOI: 10.1002/cpe.6487. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85109874243&doi=10.1002%2fcpe.6487&partnerID=40&md5=438e99b9a8ded2137620f5d8a9a5386a (zitiert auf S. 29).

[ARPL21] A. Aminifar, F. Rabbi, V. Pun, Y. Lamo. „Diversity-Aware Anonymization for Structured Health Data“. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. 2021, S. 2148–2154. DOI: 10.1109/EMBC46164.2021.9629918. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85122504835&doi=10.1109%2fEMBC46164.2021.9629918&partnerID=40&md5=1504610ed76fb64ed4ca2b6af9a71903 (zitiert auf S. 29).

[ATA18] S. Al-Kharji, Y. Tian, M. Al-Rodhaan. „A Novel (K, X)-isomorphism Method for Protecting Privacy in Weighted social Network“. In: 21st Saudi Computer Society National Computer Conference, NCC 2018. 2018. DOI: 10.1109/NCG.2018.8593107. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061480442&doi=10.1109%2fNCG.2018.8593107&partnerID=40&md5=b95cc4ca1970ef0823ad07e12f024f96 (zitiert auf S. 29).

[BJW18] S. Bazai, J. Jang-Jaccard, R. Wang. *Anonymizing k-NN classification on mapreduce*. Bd. 235. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST. Pages: 377. 2018. 364 S. DOI: 10.1007/978-3-319-90775-8_29. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047498473&doi=10.1007%2f978-3-319-90775-8_29&partnerID=40&md5=9435fd6566f8c3940954a205d2302aa9 (zitiert auf S. 26, 32).

[BK96] D. Bernstein, A. Kornhauser. „An Introduction to Map Matching for Personal Navigation Assistants“. In: (Aug. 1996) (zitiert auf S. 56, 57).

[CNV+21] G. M. Csányi, D. Nagy, R. Vági, J. P. Vadász, T. Orosz. „Challenges and Open Problems of Legal Document Anonymization“. In: *Symmetry* 13.8 (Aug. 2021). Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, S. 1490. ISSN: 2073-8994. DOI: 10.3390/sym13081490. URL: https://www.mdpi.com/2073-8994/13/8/1490 (besucht am 11. 01. 2023) (zitiert auf S. 17).

[Dud] Dudenredaktion. *Duden | anonymisieren | Rechtschreibung, Bedeutung, Definition, Herkunft*. URL: https://www.duden.de/rechtschreibung/anonymisieren (besucht am 11. 02. 2023) (zitiert auf S. 15).

[Dwo08] C. Dwork. „Differential Privacy: A Survey of Results“. In: *Theory and Applications of Models of Computation*. Hrsg. von M. Agrawal, D. Du, Z. Duan, A. Li. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, S. 1–19. ISBN: 978-3-540-79228-4. DOI: 10.1007/978-3-540-79228-4_1 (zitiert auf S. 18, 21).

[EAZS18] C. Eyupoglu, M. Aydin, A. Zaim, A. Sertbas. „An efficient big data anonymization algorithm based on chaos and perturbation techniques“. In: *Entropy* 20.5 (2018). DOI: 10.3390/e20050373. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057860878&doi=10.3390%2fe20050373&partnerID=40&md5=7394a2012aae8a6a38c37e88a184de27 (zitiert auf S. 29, 33).

[Ebe] D. Eberly. „Minimum-Area Rectangle Containing a Set of Points“. In: () (zitiert auf S. 49).

[EM19] S. H. Erfani, R. Mortazavi. „A Novel Graph-modification Technique for User Privacy-preserving on Social Networks“. In: *Journal of Telecommunications and Information Technology* 3.2019 (Okt. 2019), S. 27–38. ISSN: 15094553, 18998852. DOI: 10.26636/jtit.2019.134319. URL: https://www.itl.waw.pl/czasopisma/JTIT/2019/3/27.pdf (besucht am 22. 11. 2022) (zitiert auf S. 32).

[Fie23] A. Fieschi. *Anonymization Use Cases for Data Transfer in the Automotive Domain*. 2023 (zitiert auf S. 35).

[GL19a] T. Gao, F. Li. „PHDP: Preserving Persistent Homology in Differentially Private Graph Publications“. In: Proceedings - IEEE INFOCOM. Bd. 2019-April. 2019, S. 2242–2250. DOI: 10.1109/INFOCOM.2019.8737584. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85068229216&doi=10.1109%2fINFOCOM.2019.8737584&partnerID=40&md5=016eb888d60afe0916c71909ddb5ba25 (zitiert auf S. 28, 29).

[GL19b] T. Gao, F. Li. „Privacy-Preserving Sketching for Online Social Network Data Publication“. In: Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks workshops. Bd. 2019-June. 2019. DOI: 10.1109/SAHCN.2019.8824823. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85073008249&doi=10.1109%2fSAHCN.2019.8824823&partnerID=40&md5=0704f4031e54c44c62ce3eacb1a7ba91 (zitiert auf S. 30).

[GL19c] T. Gao, F. Li. „Sharing Social Networks Using a Novel Differentially Private Graph Model“. In: 2019 16th IEEE Annual Consumer Communications and Networking Conference, CCNC 2019. 2019. DOI: 10.1109/CCNC.2019.8651689. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063426458&doi=10.1109%2fCCNC.2019.8651689&partnerID=40&md5=70e55251b6f6510fc031aeabd0a1dd0d (zitiert auf S. 28, 29).

[GLY+17] Q. Gong, J. Luo, M. Yang, W. Ni, X.-B. Li. „Anonymizing 1:M microdata with high utility“. In: *Knowledge-Based Systems* 115 (2017), S. 15–26. DOI: 10.1016/j.knosys.2016.10.012. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84995877414&doi=10.1016%2fj.knosys.2016.10.012&partnerID=40&md5=677e34e217bac5807890ff35017f42ff (zitiert auf S. 32).

[GYG+19] N. Guo, M. Yang, Q. Gong, Z. Chen, J. Luo. „Data anonymization based on natural equivalent class“. In: Proceedings of the 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design, CSCWD 2019. 2019, S. 22–27. DOI: 10.1109/CSCWD.2019.8791905. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071489369&doi=10.1109%2fCSCWD.2019.8791905&partnerID=40&md5=aa375a46fe7abf54f3ada58b9709954d (zitiert auf S. 26, 32).

[Ham17] J. Hamm. „Minimax filter: Learning to preserve privacy from inference attacks“. In: *Journal of Machine Learning Research* 18 (2017). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85040659901&partnerID=40&md5=f2317d8e9acf49b8a69e0b438870c14d (zitiert auf S. 28).

[HRS20] D. Hemkumar, S. Ravichandra, D. Somayajulu. „Impact of prior knowledge on privacy leakage in trajectory data publishing“. In: *Engineering Science and Technology, an International Journal* 23.6 (2020), S. 1291–1300. DOI: 10.1016/j.jestch.2020.06.002. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85087212200&doi=10.1016%2fj.jestch.2020.06.002&partnerID=40&md5=75f1be8bf5f9dee8300de4497702c9d1 (zitiert auf S. 26).

[HYL+17] J. Han, J. Yu, J. Lu, H. Peng, J. Wu. *An anonymization method to improve data utility for classification*. Bd. 10581 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 71. 2017. 57 S. DOI: 10.1007/978-3-319-69471-9_5. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85034265743&doi=10.1007%2f978-3-319-69471-9_5&partnerID=40&md5=c0a68201cc2dda3449dcd1fb346fbdd3 (zitiert auf S. 32).

[JAM+21] I. Javed, F. Alharbi, T. Margaria, N. Crespi, K. Qureshi. „PETchain: A Blockchain-Based Privacy Enhancing Technology“. In: *IEEE Access* 9 (2021), S. 41129–41143. DOI: 10.1109/ACCESS.2021.3064896. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102626489&doi=10.1109%2fACCESS.2021.3064896&partnerID=40&md5=0024cbe80d1c2d8c455b48c4e15401f5 (zitiert auf S. 28).

[JHRZ22] F. Jin, W. Hua, B. Ruan, X. Zhou. „Frequency-based Randomization for Guaranteeing Differential Privacy in Spatial Trajectories“. In: Proceedings - International Conference on Data Engineering. Bd. 2022-May. 2022, S. 1727–1739. DOI: 10.1109/ICDE53745.2022.00175. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85136425113&doi=10.1109%2fICDE53745.2022.00175&partnerID=40&md5=4c2587f48f5a25d629ddcca5830012d0 (zitiert auf S. 28, 32, 33).

[KBS21] P. Kalia, D. Bansal, S. Sofat. „Privacy Preservation in Cloud Computing Using Randomized Encoding“. In: *Wireless Personal Communications* 120.4 (2021), S. 2847–2859. DOI: 10.1007/s11277-021-08588-9. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107451466&doi=10.1007%2fs11277-021-08588-9&partnerID=40&md5=8d8ccd1f93f73aaf4db29bd2e64ccd00 (zitiert auf S. 29).

[KH22] H.-A. Krebs, P. Hagenweiler. *Datenanonymisierung im Kontext von Künstlicher Intelligenz und Big Data: Grundlagen – Elementare Techniken – Anwendung*. Wiesbaden: Springer Fachmedien Wiesbaden, 2022. ISBN: 978-

3-658-37587-4 978-3-658-37588-1. DOI: 10.1007/978-3-658-37588-1. URL: https://link.springer.com/10.1007/978-3-658-37588-1 (besucht am 30. 09. 2022) (zitiert auf S. 38).

[KLKH22] M. Kim, J. Lee, H. Kwon, J. Hur. „Get off of Chain: Unveiling Dark Web Using Multilayer Bitcoin Address Clustering“. In: *IEEE Access* 10 (2022), S. 70078–70091. DOI: 10.1109/ACCESS.2022.3187210. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133780547&doi=10.1109%2fACCESS.2022.3187210&partnerID=40&md5=47c25f5fa2386588c1008c52bad23fd8 (zitiert auf S. 25).

[KP19] B. Kadhiwala, S. Patel. *A Novel k-Anonymization Approach to Prevent Insider Attack in Collaborative Social Network Data Publishing*. Bd. 11952 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 275. 2019. 255 S. DOI: 10.1007/978-3-030-36945-3_14. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076884227&doi=10.1007%2f978-3-030-36945-3_14&partnerID=40&md5=066058f72818826b296a373a4f418498 (zitiert auf S. 32).

[KSC14] S. Kim, M. K. Sung, Y. D. Chung. „A framework to preserve the privacy of electronic health data streams“. In: *Journal of Biomedical Informatics*. Special Issue on Informatics Methods in Medical Privacy 50 (1. Aug. 2014), S. 95–106. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2014.03.015. URL: https://www.sciencedirect.com/science/article/pii/S1532046414000823 (besucht am 21. 11. 2022) (zitiert auf S. 32).

[KTA+22] R. Khan, X. Tao, A. Anjum, S. Malik, S. Yu, A. Khan, W. Rehman, H. Malik. „(, m)-slicedBucket privacy model for sequential anonymization for improving privacy and utility“. In: *Transactions on Emerging Telecommunications Technologies* 33.6 (2022). DOI: 10.1002/ett.4130. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092942019&doi=10.1002%2fett.4130&partnerID=40&md5=c8afc996d413c1aa4500f20a86d04b1a (zitiert auf S. 27).

[KZD22] L. Kacha, A. Zitouni, M. Djoudi. „KAB: A new k-anonymity approach based on black hole algorithm“. In: *Journal of King Saud University - Computer and Information Sciences* 34.7 (2022), S. 4075–4088. DOI: 10.1016/j.jksuci.2021.04.014. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108556399&doi=10.1016%2fj.jksuci.2021.04.014&partnerID=40&md5=8a1878209da66815b5857d23466971d9 (zitiert auf S. 26).

[LGZ+18] H. Li, F. Guo, W. Zhang, J. Wang, J. Xing. „(a,k)-Anonymous Scheme for Privacy-Preserving Data Collection in IoT-based Healthcare Services Systems“. In: *Journal of Medical Systems* 42.3 (2018). DOI: 10.1007/s10916-

018-0896-7. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85042098134&doi=10.1007%2fs10916-018-0896-7&partnerID=40&md5=48abf561138e806063d9adc6014217da (zitiert auf S. 27, 32).

[Lit] N. Litzel. *Welche Daten produziert ein vernetztes Auto?* URL: https://www.bigdata-insider.de/welche-daten-produziert-ein-vernetztes-auto-a-1017161/ (besucht am 24. 01. 2023) (zitiert auf S. 13).

[LLV07] N. Li, T. Li, S. Venkatasubramanian. „t-Closeness: Privacy Beyond k-Anonymity and l-Diversity“. In: *2007 IEEE 23rd International Conference on Data Engineering*. 2007 IEEE 23rd International Conference on Data Engineering. ISSN: 2375-026X. Apr. 2007, S. 106–115. DOI: 10.1109/ICDE.2007.367856 (zitiert auf S. 20, 21).

[LZW21] X. Li, Y. Zhu, J. Wang. „Highly Efficient Privacy Preserving Location-Based Services with Enhanced One-Round Blind Filter“. In: *IEEE Transactions on Emerging Topics in Computing* 9.4 (2021), S. 1803–1814. DOI: 10.1109/TETC.2019.2926385. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074846390&doi=10.1109%2fTETC.2019.2926385&partnerID=40&md5=5dbe8198b676024766d0f12923c92e17 (zitiert auf S. 30).

[LZX22] B. Li, H. Zhu, M. Xie. „Releasing Differentially Private Trajectories with Optimized Data Utility“. In: *Applied Sciences (Switzerland)* 12.5 (2022). DOI: 10.3390/app12052406. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125757786&doi=10.3390%2fapp12052406&partnerID=40&md5=23239449cebd91e3b734e0bc3ba971bd (zitiert auf S. 28, 31).

[MDM22] S. Mahdavifar, F. Deldar, H. Mahdikhani. „Personalized Privacy-Preserving Publication of Trajectory Data by Generalization and Distortion of Moving Points“. In: *Journal of Network and Systems Management* 30.1 (2022). DOI: 10.1007/s10922-021-09617-5. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116850420&doi=10.1007%2fs10922-021-09617-5&partnerID=40&md5=49ce6f1990fc04623e0b8ab25e66ac51 (zitiert auf S. 31).

[MGN20] M. Mohamed, S. Ghanem, M. Nagi. „Privacy-preserving for distributed data streams: Towards l-diversity“. In: *International Arab Journal of Information Technology* 17.1 (2020), S. 52–64. DOI: 10.34028/iajit/17/1/7. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85078248649&doi=10.34028%2fiajit%2f17%2f1%2f7&partnerID=40&md5=f7c44408d9d93ad609839572256d6c9a (zitiert auf S. 31).

[MGRM19] B. Mehta, R. Gupta, U. Rao, M. Muthiyan. „A Scalable (, k)-Anonymization Approach using MapReduce for Privacy Preserving Big Data Publishing“. In: 2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019. 2019. DOI: 10.1109/ICCCNT45670.2019.8944778. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85078139154&doi=10.1109%2fICCCNT45670.2019.8944778&partnerID=40&md5=fbb213ccd2e3e63c60ccad5b9c5be5e1 (zitiert auf S. 32).

[MKGV07] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam. „L-diversity: Privacy beyond k-anonymity“. In: *ACM Transactions on Knowledge Discovery from Data* 1.1 (1. März 2007), 3–es. ISSN: 1556-4681. DOI: 10.1145/1217299.1217302. URL: https://doi.org/10.1145/1217299.1217302 (besucht am 06. 10. 2022) (zitiert auf S. 19).

[PA20a] E. Piacentino, C. Angulo. *Anonymizing Personal Images Using Generative Adversarial Networks*. Bd. 12108 LNBI. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 405. 2020. 395 S. DOI: 10.1007/978-3-030-45385-5_35. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85085165169&doi=10.1007%2f978-3-030-45385-5_35&partnerID=40&md5=24b95337b4d90f88fa997a0af1528a38 (zitiert auf S. 28).

[PA20b] E. Piacentino, C. Angulo. *Generating Fake Data Using GANs for Anonymizing Healthcare Data*. Bd. 12108 LNBI. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 417. 2020. 406 S. DOI: 10.1007/978-3-030-45385-5_36. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85085184300&doi=10.1007%2f978-3-030-45385-5_36&partnerID=40&md5=d86f7fd0e85dd73c523e732d3b6f61d9 (zitiert auf S. 28, 37).

[PKS22] V. Puri, P. Kaur, S. Sachdeva. „(k, m, t)-anonymity: Enhanced privacy for transactional data“. In: *Concurrency and Computation: Practice and Experience* 34.18 (2022). DOI: 10.1002/cpe.7020. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85127654710&doi=10.1002%2fcpe.7020&partnerID=40&md5=01c1873c8f5ee5407d136fdddf6519b2 (zitiert auf S. 31).

[PMB+21] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, D. Moher. „The PRISMA 2020 statement: an updated guideline for reporting systematic reviews“. In: *BMJ* 372 (29. März 2021). Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting, n71. ISSN: 1756-1833. DOI: 10.1136/bmj.n71. URL: https://www.bmj.com/content/372/bmj.n71 (besucht am 18. 01. 2023) (zitiert auf S. 24).

[RAD19] N. Rajesh, S. Abraham, S. S. Das. „Personalized trajectory anonymization through sensitive location points hiding“. In: *International Journal of Information Technology* 11.3 (1. Sep. 2019), S. 461–465. ISSN: 2511-2112. DOI: 10.1007/s41870-018-0261-6. URL: https://doi.org/10.1007/s41870-018-0261-6 (besucht am 23. 02. 2023) (zitiert auf S. 14, 44, 45, 47, 49).

[RMTC18] H. Rong, T. Ma, M. Tang, J. Cao. „A novel subgraph K+-isomorphism method in social network based on graph similarity detection“. In: *Soft Computing* 22.8 (2018), S. 2583–2601. DOI: 10.1007/s00500-017-2513-y. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85012914294&doi=10.1007%2fs00500-017-2513-y&partnerID=40&md5=4072f1a6f86063b949a96123e4aebdb4 (zitiert auf S. 29).

[RN22] N. Raju, V. Naresh. „Dynamic distributed KCi-slice data publishing model with multiple sensitive attributes“. In: *Concurrency and Computation: Practice and Experience* 34.21 (2022). DOI: 10.1002/cpe.7064. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85130731636&doi=10.1002%2fcpe.7064&partnerID=40&md5=9b4736461260c720578d4589ed79feaa (zitiert auf S. 27).

[SG22] A. Singh, R. Gupta. „A privacy-preserving model based on differential approach for sensitive data in cloud environment“. In: *Multimedia Tools and Applications* 81.23 (2022), S. 33127–33150. DOI: 10.1007/s11042-021-11751-w. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128266986&doi=10.1007%2fs11042-021-11751-w&partnerID=40&md5=fcd85a2f455ed3081f9e5e2aa0eb8fa3 (zitiert auf S. 28).

[SNNR22] C. Sowmyarani, L. Namya, G. Nidhi, P. Ramakanth Kumar. „Enhanced k-Anonymity model based on clustering to overcome Temporal attack in Privacy Preserving Data Publishing“. In: 2022 IEEE International Conference on Electronics, Computing and Communication Technologies, CONECCT 2022. 2022. DOI: 10.1109/CONECCT55679.2022.9865682. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138282703&doi=10.1109%2fCONECCT55679.2022.9865682&partnerID=40&md5=ac209670943c6db8731a873bba9ee260 (zitiert auf S. 26).

[SO17] Y. Sei, A. Ohsuga. „Location Anonymization with Considering Errors and Existence Probability“. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47.12 (2017), S. 3207–3218. DOI: 10.1109/TSMC.2016.2564928. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85040600139&doi=10.1109%2fTSMC.2016.2564928&partnerID=40&md5=fb686e72819ade2ce552567136f26361 (zitiert auf S. 31).

[SOTO19] Y. Sei, H. Okumura, T. Takenouchi, A. Ohsuga. „Anonymization of Sensitive Quasi-Identifiers for l-Diversity and t-Closeness“. In: *IEEE Transactions on Dependable and Secure Computing* 16.4 (2019), S. 580–593. DOI: 10.1109/TDSC.2017.2698472. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85068941270&doi=10.1109%2fTDSC.2017.2698472&partnerID=40&md5=743a3cdaaf0bcdaab3a58c18ad4ccab8 (zitiert auf S. 27, 32).

[Sup] E. D. P. Supervisor. *MISUNDERSTANDINGS RELATED TO ANONYMISATION*. European Data Protection Supervisor. URL: https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf (besucht am 10. 01. 2023) (zitiert auf S. 16).

[SVSJ19] G. Sasirekha, S. Verma, S. Sheik Faritha Begum, J. Johnsana. „An improved clustering realized relational data anonymization with optimal privacy and utility measures“. In: *International Journal of Recent Technology and Engineering* 8.1 (2019), S. 380–387. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85068878380&partnerID=40&md5=3b6bf39b1755cf7db1b4a419cdeea97c (zitiert auf S. 26, 32).

[Swe02] L. Sweeney. „k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY“. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5 (Okt. 2002), S. 557–570. ISSN: 0218-4885, 1793-6411. DOI: 10.1142/S0218488502001648. URL: https://www.worldscientific.com/doi/abs/10.1142/S0218488502001648 (besucht am 13. 01. 2023) (zitiert auf S. 17, 18).

[TCC20] M. Thouvenot, O. Curé, P. Calvez. „Knowledge graph anonymization using semantic anatomization“. In: CEUR Workshop Proceedings. Bd. 2721. 2020, S. 129–133. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096221041&partnerID=40&md5=a795009253590b6d8db0a7da2b8c9328 (zitiert auf S. 30).

[TWTH20] Y.-C. Tsai, S.-L. Wang, I.-H. Ting, T.-P. Hong. „Flexible sensitive K-anonymization on transactions“. In: *World Wide Web* 23.4 (2020), S. 2391–2406. DOI: 10.1007/s11280-020-00798-8. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084684100&doi=10.1007%2fs11280-020-00798-8&partnerID=40&md5=8ccb02a6e6586b3f00a34f0b45b50e27 (zitiert auf S. 31).

[VDW+21] J. Verdonck, K. De Boeck, M. Willocx, J. Lapon, V. Naessens. „A clustering approach to anonymize locations during dataset de-identification“. In: ACM International Conference Proceeding Series. 2021. DOI: 10.1145/3465481.3470020. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113225055&doi=10.1145%2f3465481.3470020&partnerID=40&md5=b28a34e79be778c2383b275057999cd5 (zitiert auf S. 26, 32, 33).

[Vol22] N. Vollmer. *Erwägungsgrund 26 EU Datenschutz Grundverordnung (EU-DSGVO)*. Publisher: SecureDataService. 22. Aug. 2022. URL: https://www.privacy-regulation.eu/de/erwaegungsgrund-26-DS-GVO.htm (besucht am 11. 02. 2023) (zitiert auf S. 16).

[WL18] L.-E. Wang, X. Li. „A graph-based multifold model for anonymizing data with attributes of multiple types“. In: *Computers and Security* 72 (2018), S. 122–135. DOI: 10.1016/j.cose.2017.09.003. URL: https://www.scopus.

com/inward/record.uri?eid=2-s2.0-85030689561&doi=10.1016%2fj.cose.2017.09.003&partnerID=40&md5=49d8a94e4cdf60bc6f733a094d0b9465 (zitiert auf S. 32, 33).

[WLM17] K. Ward, D. Lin, S. Madria. „MELT: Mapreduce-based efficient large-scale trajectory anonymization“. In: ACM International Conference Proceeding Series. Bd. Part F128636. 2017. DOI: 10.1145/3085504.3085581. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85025643377&doi=10.1145%2f3085504.3085581&partnerID=40&md5=5055862a35590c6c6726c424f875c619 (zitiert auf S. 31).

[WNV18] D. Wei, K. Natesan Ramamurthy, K. Varshney. „Distribution-preserving k-anonymity“. In: *Statistical Analysis and Data Mining* 11.6 (2018), S. 253–270. DOI: 10.1002/sam.11374. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85043680121&doi=10.1002%2fsam.11374&partnerID=40&md5=9e717a9ab1650eb763ba1bc2ca8abf19 (zitiert auf S. 26).

[ZGB+18] S. Zouinina, N. Grozavu, Y. Bennani, A. Lyhyaoui, N. Rogovschi. „Efficient k-Anonymization through Constrained Collaborative Clustering“. In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2018 IEEE Symposium Series on Computational Intelligence (SSCI). Nov. 2018, S. 405–411. DOI: 10.1109/SSCI.2018.8628635 (zitiert auf S. 26).

[ZLL17] S. Zhang, Q. Liu, Y. Lin. „Anonymizing popularity in online social networks with full utility“. In: *Future Generation Computer Systems* 72 (2017), S. 227–238. DOI: 10.1016/j.future.2016.05.007. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84971634640&doi=10.1016%2fj.future.2016.05.007&partnerID=40&md5=2d6f1baa3e5f81548a7d025804b5e021 (zitiert auf S. 32).

Appendix A: Dokumente der Literaturrecherche

Jan Herkommer

März 2023

References

(1) Abay, N. C.; Zhou, Y.; Kantarcioglu, M.; Thuraisingham, B.; Sweeney, L. In *Machine Learning and Knowledge Discovery in Databases*, ed. by Berlingerio, M.; Bonchi, F.; Gärtner, T.; Hurley, N.; Ifrim, G., Springer International Publishing: Cham, 2019, pp 510–526.

(2) Abbasi, A.; Mohammadi, B. *Concurrency and Computation: Practice and Experience* **2022**, *34*, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.6487, e6487.

(3) AL-Kharji, S.; Tian, Y.; Al-Rodhaan, M. In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, 2018 21st Saudi Computer Society National Computer Conference (NCC), 2018, pp 1–6.

(4) Aleroud, A.; Shariah, M.; Malkawi, R. In *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2022 IEEE International Conference on Cyber Security and Resilience (CSR), 2022, pp 357–363.

(5) Aljably, R. In *Big Data and Security*, ed. by Tian, Y.; Ma, T.; Khan, M. K., Springer: Singapore, 2021, pp 142–152.

(6) Alnemari, A.; Arodi, S.; Sosa, V.; Pandey, S.; Romanowski, C.; Raj, R.; Mishra, S. *IFIP Advances in Information and Communication Technology* **2018**, *542*, ISBN: 9783030045364, 113–125.

(7) Aminifar, A.; Rabbi, F.; Pun, V. K. I.; Lamo, Y. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); ISSN: 2694-0604, 2021, pp 2148–2154.

(8) An, S.; Li, Y.; Wang, T.; Jin, Y. In *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*, 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC), 2018, pp 132–137.

(9) Ashoka, K.; Poornima, B. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp 970–976.

(10) Ashoka, K.; Poornima, B. In *Smart Computing and Informatics*, ed. by Satapathy, S. C.; Bhateja, V.; Das, S., Springer: Singapore, 2018, pp 593–601.

(11) Attaullah, H.; Anjum, A.; Kanwal, T.; Malik, S.; Asheralieva, A.; Malik, H.; Zoha, A.; Arshad, K.; Imran, M. *Sensors* **2021**, *21*, DOI: `10.3390/s21144933`.

(12) Bazai, S. U.; Jang-Jaccard, J.; Wang, R. In *Mobile Networks and Management*, ed. by Hu, J.; Khalil, I.; Tari, Z.; Wen, S., Springer International Publishing: Cham, 2018, pp 364–377.

(13) Cohen-Hadria, A.; Cartwright, M.; McFee, B.; Bello, J. P. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP); ISSN: 1551-2541, 2019, pp 1–6.

(14) Erfani, S. H.; Mortazavi, R. *Journal of Telecommunications and Information Technology* **2019**, *3*, 27–38.

(15) Eyupoglu, C.; Aydin, M.; Zaim, A.; Sertbas, A. *Entropy* **2018**, *20*, DOI: `10.3390/e20050373`.

(16) Gangarde, R.; Sharma, A.; Pawar, A.; Joshi, R.; Gonge, S. *Electronics (Switzerland)* **2021**, *10*, DOI: `10.3390/electronics10222877`.

(17) Gao, T.; Li, F. In Proceedings - IEEE INFOCOM; ISSN: 0743-166X, 2019; Vol. 2019-April, pp 2242–2250.

(18) Gao, T.; Li, F. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON); ISSN: 2155-5494, 2019, pp 1–9.

(19) Gao, T.; Li, F. In *2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, 2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC); ISSN: 2331-9860, 2019, pp 1–4.

(20) Gong, Q.; Luo, J.; Yang, M.; Ni, W.; Li, X.-B. *Knowledge-Based Systems* **2017**, *115*, 15–26.

(21) Guo, N.; Yang, M.; Gong, Q.; Chen, Z.; Luo, J. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2019, pp 22–27.

(22) Hamm, J. Minimax Filter: Learning to Preserve Privacy from Inference Attacks, 2017.

(23) Han, J.; Yu, J.; Lu, J.; Peng, H.; Wu, J. In *Cyberspace Safety and Security*, ed. by Wen, S.; Wu, W.; Castiglione, A., Springer International Publishing: Cham, 2017, pp 57–71.

(24) Hemkumar, D.; Ravichandra, S.; Somayajulu, D. *Engineering Science and Technology, an International Journal* **2020**, *23*, 1291–1300.

(25) Imtiaz, S.; Tania, Z. N.; Nazeer Chaudhry, H.; Arsalan, M.; Sadre, R.; Vlassov, V. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, 2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), 2021, pp 1592–1602.

(26) Javed, I.; Alharbi, F.; Margaria, T.; Crespi, N.; Qureshi, K. *IEEE Access* **2021**, *9*, 41129–41143.

(27) Jayapradha, J.; Prakash, M. *Soft Computing* **2022**, *26*, 13019–13036.

(28) Jha, N.; Favale, T.; Vassio, L.; Trevisan, M.; Mellia, M. In *2020 IEEE International Conference on Big Data (Big Data)*, 2020 IEEE International Conference on Big Data (Big Data), 2020, pp 3996–4005.

(29) Jin, F.; Hua, W.; Ruan, B.; Zhou, X. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 2022 IEEE 38th International Conference on Data Engineering (ICDE); ISSN: 2375-026X, 2022, pp 1727–1739.

(30) Kacha, L.; Zitouni, A.; Djoudi, M. *Journal of King Saud University - Computer and Information Sciences* **2022**, *34*, 4075–4088.

(31) Kadhiwala, B.; Patel, S. J. In *Information Systems Security*, ed. by Garg, D.; Kumar, N. V. N.; Shyamasundar, R. K., Springer International Publishing: Cham, 2019, pp 255–275.

(32) Kalia, P.; Bansal, D.; Sofat, S. *Wireless Personal Communications* **2021**, *120*, 2847–2859.

(33) Khan, R.; Tao, X.; Anjum, A.; Malik, S.; Yu, S.; Khan, A.; Rehman, W.; Malik, H. *Transactions on Emerging Telecommunications Technologies* **2022**, *33*, DOI: `10.1002/ett.4130`.

(34) Khan, R.; Tao, X.; Anjum, A.; Sajjad, H.; Malik, S.; Khan, A.; Amiri, F. *Wireless Communications and Mobile Computing* **2020**, *2020*, DOI: `10.1155/2020/8416823`.

(35) Kim, S.; Sung, M. K.; Chung, Y. D. *Journal of Biomedical Informatics* **2014**, *50*, 95–106.

(36) Li, B.; Zhu, H.; Xie, M. *Applied Sciences (Switzerland)* **2022**, *12*, DOI: `10.3390/app12052406`.

(37) Li, H.; Guo, F.; Zhang, W.; Wang, J.; Xing, J. *Journal of Medical Systems* **2018**, *42*, 56.

(38) Li, X.; Zhu, Y.; Wang, J. *IEEE Transactions on Emerging Topics in Computing* **2021**, *9*, Conference Name: IEEE Transactions on Emerging Topics in Computing, 1803–1814.

(39) Liu, P.; Wang, H.; Lin, S.; Li, X. *Journal of Internet Technology* **2018**, *19*, Number: 1, 247–256.

(40) Liu, W.; Pei, M.; Cheng, C.; She, W.; Wu, C. Q. In *Proceedings of the 2nd International Conference on Healthcare Science and Engineering*, ed. by Wu, C. Q.; Chyu, M.-C.; Lloret, J.; Li, X., Springer: Singapore, 2019, pp 115–128.

(41) Mahanan, W.; Chaovalitwongse, W. A.; Natwichai, J. *Service Oriented Computing and Applications* **2020**, *14*, 89–100.

(42) Mahdavifar, S.; Deldar, F.; Mahdikhani, H. *Journal of Network and Systems Management* **2021**, *30*, 10.

(43) Majeed, A.; Lee, S. *Applied Intelligence* **2020**, *50*, 2555–2574.

(44) Medková, J. In *2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS)*, 2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS), 2020, pp 1–6.

(45) Mehta, B. B.; Gupta, R.; Rao, U. P.; Muthiyan, M. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp 1–6.

(46) Mohamed, M.; Ghanem, S.; Nagi, M. *The International Arab Journal of Information Technology* **2020**, 52–64.

(47) Al-Obaidi, S.; Al-Khafaji, H.; Abhayaratne, C. *IEEE Access* **2020**, *8*, Conference Name: IEEE Access, 213806–213824.

(48) Ouazzani, Z. E.; Bakkali, H. E. In *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*, 2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), 2017, pp 1–6.

(49) Piacentino, E.; Angulo, C. In *Bioinformatics and Biomedical Engineering*, ed. by Rojas, I.; Valenzuela, O.; Rojas, F.; Herrera, L. J.; Ortuño, F., Springer International Publishing: Cham, 2020, pp 395–405.

(50) Piacentino, E.; Angulo, C. In *Bioinformatics and Biomedical Engineering*, ed. by Rojas, I.; Valenzuela, O.; Rojas, F.; Herrera, L. J.; Ortuño, F., Springer International Publishing: Cham, 2020, pp 406–417.

(51) Podlesny, N. J.; Kayem, A. V.; Meinel, C. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 2019, pp 556–563.

(52) Puri, V.; Kaur, P.; Sachdeva, S. *Concurrency and Computation: Practice and Experience* **2022**, *34*, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.7020, e7020.

(53) Raju, N. V. S. L.; Naresh, V. S. *Concurrency and Computation: Practice and Experience* **2022**, *34*, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.7064, e7064.

(54) Riyazuddin, M.; Balaram, V. V. S. S. S. In *Proceedings of the First International Conference on Computational Intelligence and Informatics*, ed. by Satapathy, S. C.; Prasad, V. K.; Rani, B. P.; Udgata, S. K.; Raju, K. S., Springer: Singapore, 2017, pp 603–614.

(55) Rong, H.; Ma, T.; Tang, M.; Cao, J. *Soft Computing* **2018**, *22*, 2583–2601.

(56) Rousseau, F.; Casas-Roma, J.; Vazirgiannis, M. *Knowledge and Information Systems* **2018**, *54*, 315–343.

(57) Sasirekha, G.; et al. *International Journal of Recent Technology and Engineering (IJRTE).*

(58) Sei, Y.; Ohsuga, A. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **2017**, *47*, Conference Name: IEEE Transactions on Systems, Man, and Cybernetics: Systems, 3207–3218.

(59) Sei, Y.; Okumura, H.; Takenouchi, T.; Ohsuga, A. *IEEE Transactions on Dependable and Secure Computing* **2019**, *16*, Conference Name: IEEE Transactions on Dependable and Secure Computing, 580–593.

(60) Singh, A. K.; Gupta, R. *Multimedia Tools and Applications* **2022**, *81*, 33127–33150.

(61) Sowmyarani, C. N.; Namya, L. G.; Nidhi, G. K.; Ramakanth Kumar, P. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT); ISSN: 2766-2101, 2022, pp 1–6.

(62) Stanciu, V.-D.; van Steen, M.; Dobre, C.; Peter, A. In *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, Association for Computing Machinery: New York, NY, USA, 2021, pp 376–385.

(63) Sun, G.; Chang, V.; Ramachandran, M.; Sun, Z.; Li, G.; Yu, H.; Liao, D. *Journal of Network and Computer Applications* **2017**, *89*, 3–13.

(64) Thouvenot, M.; Curé, O.; Calvez, P. In *2020 IEEE International Conference on Big Data (Big Data)*, 2020 IEEE International Conference on Big Data (Big Data), 2020, pp 4065–4074.

(65) Tsai, Y.-C.; Wang, S.-L.; Ting, I.-H.; Hong, T.-P. In *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*, 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC), 2018, pp 201–206.

(66) Tsai, Y.-C.; Wang, S.-L.; Ting, I.-H.; Hong, T.-P. *World Wide Web* **2020**, *23*, 2391–2406.

(67) Verdonck, J.; De Boeck, K.; Willocx, M.; Lapon, J.; Naessens, V. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, Association for Computing Machinery: New York, NY, USA, 2021, pp 1–10.

(68) Virupaksha, S.; Dondeti, V. *Peer-to-Peer Networking and Applications* **2021**, *14*, 1608–1628.

(69) Wang, L.-E.; Li, X. *Computers & Security* **2018**, *72*, 122–135.

(70) Ward, K.; Lin, D.; Madria, S. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, Association for Computing Machinery: New York, NY, USA, 2017, pp 1–6.

(71) Wei, D.; Natesan Ramamurthy, K.; Varshney, K. *Statistical Analysis and Data Mining* **2018**, *11*, 253–270.

(72) Widodo; Budiardjo, E.; Wibowo, W. *Information (Switzerland)* **2019**, *10*, DOI: `10.3390/info10120362`.

(73) Yamaç, M.; Ahishali, M.; Passalis, N.; Raitoharju, J.; Sankur, B.; Gabbouj, M. In *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019 27th European Signal Processing Conference (EUSIPCO); ISSN: 2076-1465, 2019, pp 1–5.

(74) Zhang, J.; Li, H.; Liu, X.; Luo, Y.; Chen, F.; Wang, H.; Chang, L. *IEEE Transactions on Dependable and Secure Computing* **2017**, *14*, Conference Name: IEEE Transactions on Dependable and Secure Computing, 507–520.

(75) Zhang, S.; Liu, Q.; Lin, Y. *Future Generation Computer Systems* **2017**, *72*, 227–238.

(76) Zouinina, S.; Grozavu, N.; Bennani, Y.; Lyhyaoui, A.; Rogovschi, N. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp 405–411.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift