Forschungsprojekt EIDI

Effektive Information nach digitalem Identitätsdiebstahl

DELIVERABLE 2.6

Datenschutzrelevante Merkmale digitaler Identitäten

ausgearbeitet von

Bud P. Bruegger

ULD SCHLESWIG-HOLSTEIN

Kiel, im März 2018

Projektpartner

RHEINISCHE FRIEDERICH-WILLHELMS UNIVERISÄT BONN

Informatik 4

Friedrich-Ebert-Allee 144
53113 Bonn

Prof. Dr. Michael Meier
0228 7354249
mm@cs.uni-bonn.de

UNABHÄNGIGES LANDESZENTRUM FÜR DATENSCHUTZ SCHLESWIG-HOLSTEIN

Holstenstr. 98
24103 Kiel

Harald Zwingelberg
0431 98812222
uld6@datenschutzzentrum.de

LEIBNITZ-INSTITUT FÜR INFORMATIONSINFRASTRUKTUR GMBH KARLSRUHE

Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldhafen

Prof. Dr. Franziska Boehm
07247 808555
franziska.boehm@kit.edu

UNIVERSITÄT DUISBURG-ESSEN

Allgemeine Psychologie: Kognition

Forsthausweg 2
47057 Duisburg

Prof. Dr. Matthias Brand
0203 3792541
matthias.brand@uni-due.de

XING AG
Dammtorstraße 30
20354 Hamburg

1. Einführung

Das vorliegende Deliverable befasst sich mit datenschutzrelevanten Merkmalen digitaler Identitäten. In einem ersten Schritt wird der Begriff der digitalen Identität definiert. In einem zweiten Schritt wird dann die Situation von EIDI näher betrachtet, um die für das Projekt relevanten Merkmale identifizieren zu können. In weiteren Schritten werden dann die identifizierten Gruppen von Merkmalen detailliert beschrieben. Diese umfassen die folgenden:

- Faktoren, die die Schwere des Risikos für Betroffene beeinflussen.
- Risiken, die durch EIDI beeinflusst werden können.
- Risiken, die nicht mehr von EIDI beeinflusst werden können.
- Durch EIDI selbst verursachte Risiken.

Zusätzlich zu dieser systematischen Erfassung digitaler Identitäten und deren Merkmalen wird eine Klassifizierung von Anbietern von digitalen Identitäten vorgeschlagen und auf die Frage eingegangen, unter welchen Umständen eine Verkettung von Datensätzen in EIDI notwendig ist. Das Deliverable endet dann mit einem Fazit.

2. Definition des Begriffs „digitale Identität“

„Digitale Identität“ ist ein zentrales Konzept im EIDI Projekt. Um ein einheitliches Verständnis des Begriffs im Konsortium zu erreichen und da die Definition auch den Umfang der Arbeit beeinflusst, hat das Projekt am 23.11.2017 einen internen Workshop zum Thema

in Berlin organisiert. Diese Sektion stellt weitere Hintergrundinformation zum Begriff der digitalen Identität zusammen. Insbesondere wird eine Übersicht von verfügbaren Definitionen des Begriffs gegeben und die im Workshop abgestimmte Definition festgehalten. Des Weitern werden einige relevante Eigenschaften von digitalen Identitäten diskutiert.

2.1 Verfügbare Definitionen des Begriffs „digitale Identität“

Es existiert eine Vielfalt von verschiedenen Definitionen von digitaler Identität. Die Situation wurde treffend vom NIST beschrieben: „Digital identity is the online persona of a subject, and a single definition is widely debated internationally.”[1]

Insbesondere nehmen die diversen Anwendungsgebiete verschiedene Blickwinkel ein und stellen andere Anforderungen an das Konzept der digitalen Identität. Ohne Anspruch auf Vollständigkeit sind einige dieser Anwendungsgebiete (i) Zugangskontrolle, (ii) soziale Medien und (iii) Datenschutz.

(i) Im Gebiet der Zugangskontrolle steht im Vordergrund, wie man über ein Netzwerk seine Identität nachweisen kann. Wichtig dabei ist typisch ein geheimes Element, das nur dem legitimen Halter der Identität bekannt oder zugänglich ist. Oft spricht man in diesem Zusammenhang auch von Credentials, Tokens, Zertifikaten, Assertions oder Karten (eIDs). Das wichtigste (und z.T. einzige) Attribut einer Person ist dann eine eindeutige Kennung wie z.B. ein Benutzername oder eine Kunden- oder Steuernummer. Diese werden typischerweise vom „Ausgeber der Identität“ (identity provider) zugewiesen oder verwaltet. Digitale Identitäten in diesem Sinne können optional auch wenige zusätzliche Attribute einer Person beinhalten. Typischerweise ist dann der Ausgeber eine Autorität oder zumindest ein vertrauenswürdiger Dritter derjenige, der die Richtigkeit dieser Attribute gewährleistet (Zertifizierung oder Assertion).

(ii) Im Gebiet der sozialen Netzwerke werden digitale Identitäten durch die Erstellung von Account eines Benutzers kreiert. Diese Identitäten haben zwei Aspekte: Einerseits unterstützen diese Identitäten über Technologien des föderierten Identitätsmanagements die Zugangskontrolle für eine offene Zahl von Dienstleistungen im Internet, andererseits kann man die über den Account dargestellten Daten als eine Projektion der physischen Person ins Internet (in den Cyberspace) sehen. Für Zugangskontrolle wird vorwiegend eine E-Mail-Adresse verwendet; auf dem Account sammelt sich aber eine Vielfalt von anderen Attributen

[1] Paul A. Grassi, Michael E. Garcia, James L. Fenton: NIST Special Publication 800-63-3, Digital Identity Guidelines, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-3.pdf, June 2017, Executive Summary, page iv.

der Person an. Diese Attribute werden aber nicht durch den Betreiber der sozialen Plattform garantiert. Die Vertrauenswürdigkeit solcher Attribute wird durch den Zeitraum der sichtbaren Aktivität und durch Verlinkungen mit Identitäten von andern sozialen Netzwerken abgeleitet. Zum Beispiel ist eine gerade eben erstellte Identität wenig vertrauenswürdig. Dagegen wir eine lang existierende Identität mit guter Verlinkung zu schon vertrauten Identitäten typischerweise als vertrauenswürdig eingestuft. Die Sichtweise der Identitäten von sozialen Netzwerken erweitert damit drastisch die möglichen Attribute einer Person und verzichtet auf den Anspruch, dass Attribute garantiert werden müssen.

(iii) Die Sichtweise des Datenschutzes erweitert den Umfang einer digitalen Identität noch weiter, indem dieser ohne Einschränkungen alle möglichen Attribute einer Person und ihrer Aktivitäten in die Identität einbezieht.[2]

Insbesondere könnte man auf dieser Basis digitale Identitäten wie folgt definieren[3]:

> *Eine **digitale Identität** ist die Gesamtheit von technisch abgebildeten Attributen (Daten), die einer Person zugeordnet werden können und damit die Person in der digitalen Welt repräsentieren.*

Hansen und Meissner verwenden auch den Begriff einer „Teilidentität“[4]. Für den EIDI Workshop wurde daraus die folgende Definition abgeleitet:

> *„Eine **digitale Teilidentität** ist ein Datensatz, der eine Teilmenge von Aspekten einer Person beschreibt, die in einem bestimmten Kontext und für einen bestimmten Zweck von Interesse sind. Eine Teilidentität ist eine Untermenge der digitalen Identität.“*

Dieser Sachverhalt ist z.B. in Abbildung 1 dargestellt. (Quelle: Hansen und Meissner , Verkettung digitaler Identitäten, Abbildung 1, Seite 23).

[2] Siehe z.B. Marit Hansen und Sebastian Meissner (Editoren), Verkettung digitaler Identitäten, Oktober 2007, https://www.datenschutzzentrum.de/uploads/projekte/verkettung/2007-uld-tud-verkettung-digitaler-identitaeten-bmbf1.pdf, 2.2.3Digitale Identitäten, Seiten 22-26.
[3] Diese Definitionen von digitaler Identität und Teilidentität wurde von ULD spezifisch als Vorbereitung des EIDI Workshops über digitale Identitäten entwickelt.
[4] Seite 22, siehe Fussnote 2.

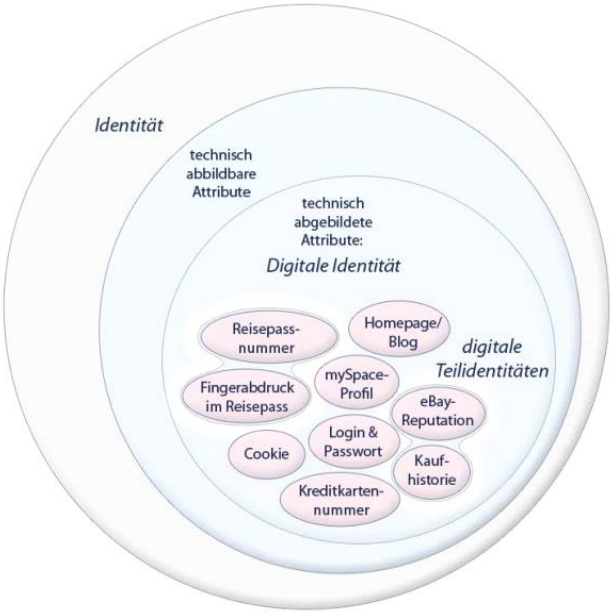

Abbildung 1: Identität, digitale Identität und digitale Teilidentität nach Hansen und Meissner.

Die oben vorgeschlagene Definition von „digitaler Identität" wirft die Frage auf, wann genau denn ein Attribut einer Person zugeordnet werden kann. Die Antwort ist eng mit der Weise verknüpft mit der wir die Welt wahrnehmen und konzeptualisieren (also über die Welt nachdenken).

Unsere Sinne, wie auch technische Sensoren, nehmen die Welt als ein Kontinuum vom Werten wahr. Um überhaupt von einer Person sprechen zu können, müssen in diesem Kontinuum zuerst diskrete Entitäten (Objekte) abstrahiert werden. Gemäß Piaget lernen wir dies in den ersten zwei Lebensjahren[5]:

> *"Children learn that they are separate from the environment. They can think about aspects of the environment, even though these may be outside the reach of the child's senses. In this stage, according to Piaget, the* ***development of object permanence*** *is one of the most important accomplishments. Object permanence is a child's understanding that objects continue to exist even though he or she cannot see or hear them. Peek-a-boo is a good test for that.* ***By the end of the sensorimotor period, children develop a permanent sense of self and object.****"(Hervorhebung durch den Autor)*

[5] Wikipedia, Piaget's theory of cognitive development, 1.2 Sensorimotor stage, https://en.wikipedia.org/wiki/Piaget%27s_theory_of_cognitive_development#Sensorimotor_stage, in der Version vom 22.3.2018.

Wenn einmal diskrete Entitäten, inklusive Personen, abstrahiert sind, was ist dann die Identität einer Entität? Wikipedia sagt dazu folgendes[6]:

> ***„Identität** (lateinisch īdem ‚derselbe', ĭdem ‚dasselbe') **ist die Gesamtheit, die eine Entität**, einen Gegenstand oder ein Objekt **kennzeichnet** und als Individuum von allen anderen unterscheidenden Eigentümlichkeiten beschreibt."(Hervorhebung durch den Autor)*

„Gesamtheit" macht deutlich, dass alle möglichen Aspekte einbezogen werden müssen und keine ausgelassen werden können. Ohne Anspruch auf Vollständigkeit umfassen die Aspekte, die einer Person zugeordnet werden können, die Folgenden:

Physische Aspekte einer Person umfassen z.B. **Eigenschaften** wie Position und Bewegung, äußerliche Merkmale des Körpers inklusive Biometrie und innere Merkmale des Körpers, wie z.B. der Gesundheitszustand. Neben Eigenschaften umfassen sie auch **Aktionen und Tätigkeiten** und physische (motorische) **Fähigkeiten** in Gebieten wie Handwerk oder Kunst.

Emotionale Aspekte einer Person umfassen z.B. Stimmung, Laune, Gemütszustand, Charakter, erlittene Traumen und Verhaltensmuster. Diese können durchaus auch digital erfasst werden, z.B. durch Analyse des Gesichtsausdrucks oder sprachlichen Äußerungen der Person. In der Informatik befasst sich z.B. Sentiment Analysis[7] mit dem Thema.

Mentale Aspekte einer Person umfassen z.B. Intelligenz, Bildung, Wissen, Fähigkeiten (wie z.B. Abstraktionsvermögen), Meinungen, Einstellungen, Überzeugungen, Kultur, Weltbild, Denkmodelle (Schulen), Interessen, Zu- und Abneigungen, Tabus, etc.

Spirituelle Aspekte umfassen z.B. Glaube und Konfession.

Die Gesamtheit umfasst über die Eigenschaften der isolierten Entität auch Beziehungen zu andern Entitäten. Dies beinhaltet z.B. die Folgenden:

Beziehungen zu Objekten wie z.B. Eigentum und Besitz, Kontrolle, Kreation und Urheberschaft.

Beziehungen zu Organisationen umfassen z.B. Anstellungs- und Vertragsverhältnisse, Repräsentation und Vertretung, Mitgliedschaft und Unterstützung.

Beziehungen zu andern Personen umfassen z.B. Verwandtschaft (biologisch), Partnerschaft, Freundschaft, Feindschaft (emotional), Bekanntschaft, Kollegenschaft, Nachbarschaft (gemeinsame Aktivität), Wohngemeinschaft, Mitbesitz (gemeinsames Objekt).

---

[6] Wikipedia, Identität, https://de.wikipedia.org/wiki/Identit%C3%A4t, in der Version vom 21.11.2017.

[7] Wikipedia, Sentiment analysis, https://en.wikipedia.org/wiki/Sentiment_analysis, in der Version vom 22.3.2018.

Mit dieser ausführlichen Beschreibung von drei verschiedenen Verständnissen von „digitaler Identität“ stellt sich nun die Frage, welche Definition die geeignetste für das EIDI Projekt ist.

2.2 Begriffsdefinition für das EIDI-Projekt

Da der Dateninhalt in den Breaches, die EIDI untersucht, nicht voraussehbar ist, ist es empfehlenswert, dass EIDI die breitestmögliche Definition von digitaler Identität verwendet. Namentlich ist das diejenige, die schon im Datenschutz üblich ist. Die anderen Definitionen des Konzepts sind dann Untermengen und Teilaspekte, die durchaus eingeschlossen bleiben.

Im oben genannten Workshop haben sich die Projektpartner auf die folgenden Definitionen für analoge und digitale Identität geeinigt:

> *Die **analoge Identität** ist die Menge aller veränderlichen und unveränderlichen Merkmale einer Person. Ein alleinstehendes Merkmal muss nicht personenbeziehbar sein. Die Kombination alleinstehender nicht personenbeziehbarer Merkmale zu einer Teilmenge kann zur Personenbeziehbarkeit führen.*
>
> *Die **digitale Identität** ist die Menge aller technisch abbildbaren Merkmale einer analogen Identität. Teilmengen dieser Merkmale definieren digitale **Teilidentitäten**. In einem bestimmten Kontext verwendete digitale Teilidentitäten sind dabei selbst wieder Merkmal der digitalen Identität.*

In beiden Fällen sind die Identitätsinhaber immer natürliche Personen.

2.3 Eigenschaften digitaler Identitäten

Daten repräsentieren verschiedene Aspekte von Identitäten. Um Daten richtig interpretieren zu können, scheint ein kurzer Exkurs über ihre Herkunft und Eigenschaften hilfreich.

Daten können in verschiedenen Arten erhoben werden. Eine der grundlegenden Unterscheidungen ist die Herkunft der Daten. Daten können:

- von der betroffenen Person angegeben werden oder
- von einer dritten Person über die betroffene Person gesammelt werden.

Ein Beispiel für den ersten Fall ist das Ausfüllen eines Formulars durch den Betroffenen; im zweiten Fall beobachtet ein Dritter die betroffene Person und erfasst dadurch Daten. Videoüberwachung ist ein typisches Beispiel.

Daten können:

- direkt digital anfallen oder

- analog erfasst und dann digitalisiert werden.

Offensichtlich führt die letzte Möglichkeit zusätzliche Arbeitsschritte und damit zusätzliche Fehlerquellen ein.

Ein Datensatz kann von einer einzigen Datenerhebungsaktivität stammen oder von verschiedenen Quellen kombiniert werden.

Ein Datensatz kann entweder direkt im Zustand seiner Erhebung vorliegen (Rohdaten) oder durch Analyse und Interpretation von Rohdaten erzeugt werden.

Daten sind meist technische Abbildungen von Aspekten der Wirklichkeit. Die Beziehung von Datum und Wirklichkeit ist aber meist sehr komplex. Das Folgende ist eine unvollständige Liste von relevanten Faktoren:

- Der **Kontext**, in dem ein Datum gültig ist. (z.B. Schuhgröße vs. Skischuhgröße)
- **Interpretation,** die z.B. auf **Annahmen** und Voreingenommenheit beruhen können.
- **Meinungen** und **Behauptungen**, vor allem wenn Daten von Dritten über Personen gesammelt werden.
- **Fehler**, z.B. in der Digitalisierung von analogen Grundlagen. (z.B. OCR)
- **Ungenauigkeit, Unvollständigkeit**.
- Begrenzte **zeitliche Gültigkeit**. (z.B. veraltete Daten).

Oft vergessen werden implizite Daten und Metadaten, die nicht bewusst kreiert oder versandt werden. Sie fallen z.B. bei Kommunikation oder Eingaben an. Einige Beispiele sollen dies illustrieren:

- Neben den explizit eingegebenen Daten können oft auch **Eigenschaften des Eingabegeräts** als Daten anfallen. Beispiele umfassen Browser Fingerprints, Tracking Cookies, sowie MAC und IP Adressen.
- **Sprachliche Eingaben** enthalten oft Zusatzinformationen durch Dialekt, Wortwahl und Grammatik.
- **Verbale Eingaben** übertragen neben dem sprachlichen Inhalt auch den Ton der Stimme (und damit die Stimmung). Durch die Stimme kontrollierte Eingaben erfassen neben der Stimme auch andere Geräusche, die Daten z.B. durch Tätigkeiten des Betroffenen erzeugt werden können.
- Das **Timing** von Eingaben kann zur Ableitung zusätzlicher Daten genutzt werden. Durch die Tageszeit der Eingabe können z.B. Schlüsse über den Aufenthaltsort ziehen (Arbeitsort vs. Schlafort); im Kleinen ist das relative Timing von Tastenanschlägen eine „behavioral biometrics“ und kann z.B. zur Identifikation des Eingebenden genutzt werden.

Die Möglichkeit von abgeleiteten, impliziten Daten wie Identifikation, Aufenthaltsort, Bildungsstand, kulturelle oder nationale Angehörigkeit oder Stimmung sollten also im Auge behalten werden.

3. Datenschutzrelevante Merkmale in EIDI

In dieser Sektion wird die Situation von EIDI näher betrachtet, um die für das Projekt relevanten Merkmale identifizieren zu können. Dazu wird EIDI im Kontext eines Daten-Breaches betrachtet (siehe Abbildung 2).

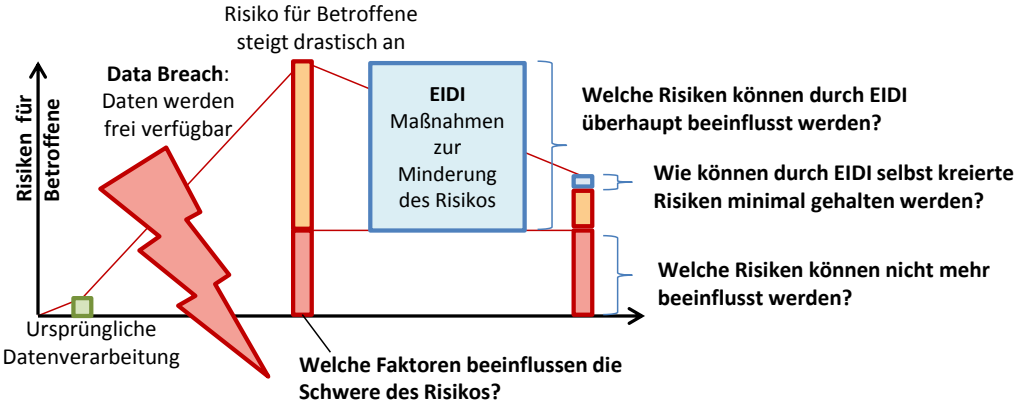

Abbildung 2: EIDI im Kontext eines Daten-Breaches.

Die Abbildung stellt auf der vertikalen Achse das Risiko für die Rechte und Freiheiten der betroffenen Personen dar. Die horizontale Achse stellt den zeitlichen Verlauf dar. Der grüne Balken links stellt das begrenzte Risiko dar, dass durch die ursprüngliche Datenverarbeitung entsteht.

Ein Daten-Breach umgeht dann die Vertraulichkeitsmaßnahmen der ursprünglichen Verarbeitung und macht Merkmale digitaler Identitäten in verschiedenen Datensenken frei verfügbar. Der Zusammenbruch der Vertraulichkeit hat einen drastischen Anstieg des Risikos für die betroffenen Personen zur Folge. Dies ist durch den rot/orangen Balken dargestellt.

Das EIDI Projekt hat zum Ziel, das Risiko für die Betroffenen nach einem Daten-Breach zu senken. Dies kann z.B. durch Warnung der Betroffenen erfolgen. EIDI implementiert deshalb Maßnahmen zur Senkung des Risikos. Es gibt einige vom Breach resultierende Risiken,

die von EIDI nicht mehr beeinflusst werde können. Diese sind durch einen roten Balken dargestellt. Die Risiken, die von EIDI beeinflusst werden können, sind in oranger Farbe dargestellt.

Nach der Behandlung der Risiken durch EIDI Maßnahmen verbleiben Restrisiken, die wiederum als Balken dargestellt sind. Der rote Teil sind unbeeinflussbare Restrisiken, die verbleiben. Der orange Teil sind beinflussbare Risiken, die deutlich reduziert sind, aber nicht ganz eliminiert werden können. Die Datenverarbeitung durch EIDI, wie jede Datenverarbeitung, führt zu zusätzlichen Risiken für die Betroffenen. Diese sind in blau dargestellt und müssen so minimal wie möglich gehalten werden.

Basierend auf dieser Ausgangslage können nun die datenschutzrelevanten Merkmale identifiziert werden, die in diesem Deliverable näher untersucht werden sollen. Diese sind:

- Die Faktoren, welche die Schwere des Risikos in einer geleakten digitalen Identität bestimmen.
- Risiken, die von EIDI beeinflusst werden können.
- Risiken, die nicht mehr durch EIDI beeinflusst werden können.
- Durch EIDI selbst verursachte Risiken und eine Vorgehensweise zu deren Minimierung.

Diese werden im den folgenden Abschnitten analysiert.

4. Faktoren zur Einschätzung der Schwere des Risikos

Im Folgenden werden die Faktoren näher analysiert, die das Risiko für Betroffene beeinflussen. Das daraus resultierende bessere Verständnis des Risikobegriffs dient dazu, die risikomindernden Maßnahmen von EIDI zu entwerfen.

4.1 Stärke des Personenbezugs

Die Stärke des Personenbezugs ist ein wichtiger Faktor zur Einschätzung des Risikos, die frei verfügbare personenbezogene Daten für die Rechte und Freiheiten von betroffenen Personen darstellen. Insbesondere sinkt das Risiko mit absteigender Wahrscheinlichkeit einer vollen Identifikation und mit steigendem Aufwand, der für eine solche Identifikation notwendig ist. Dies ist auch der Grund, warum die DS-GVO eine frühzeitige Pseudonymisie-

rung vorschreibt und für effektiv anonymisierte Daten keine Schutzmaßnahmen mehr verlangt.

Es ist zwar nicht möglich, die Stärke des Personenbezugs in den verfügbaren Leaks zu verändern, EIDI kann diesen Faktor aber in seiner eignen Verarbeitung kontrollieren, um das durch EIDI erzeugte Risiko zu minimieren. Insbesondere kann EIDI dieses Risiko stark eingrenzen, indem Datenelemente mit hohem Identifizierungspotential so früh wie möglich pseudonymisiert oder gar gelöscht werden.

Die europäische Datenschutz-Grundverordnung definiert personenbezogene Daten als „alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (im Folgenden „betroffene Person“) beziehen“[8].

Der Unterschied zwischen identifiziert und identifizierbar ist in Abbildung 3 dargestellt. Der linke Datensatz (Teilidentität) erlaubt es, die betroffene Person direkt zu ermitteln; der rechte Datensatz erlaubt dies zwar nicht, aber die betroffene Person kann indirekt, über Verkettung mit andern Datensätzen ermittelt werden. Beispielsweise ist kann eine Teilidentität ein Pseudonym einer Person beinhalten. Im Datenschutz werden Pseudonyme so gewählt, dass sie die direkte Identifizierung der betroffenen Person verhindern. Die Person bleibt aber identifizierbar, da eine Verkettung mit einer anderen Teilidentität, die ihrerseits die Person identifiziert, immer noch möglich ist.

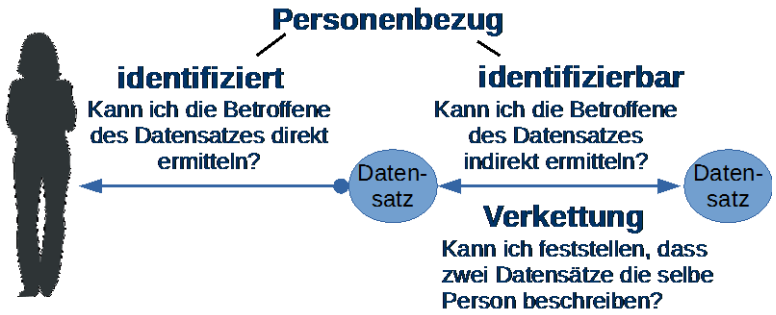

Abbildung 3: Personenbezug, Identifizirung, Identifizierbarkeit und Verkettung.

In Abbildung 4 ist dargestellt, wie Verkettung über mehrere Schritte erfolgen kann. Offensichtlich steigt das Risiko der Identifikation mit zunehmender Menge von Daten, die über eine betroffene Person zusammengestellt werden können.

[8] Datenschutz-Grundverordnung, Artikel 4, Paragraph 1, http://eur-lex.europa.eu/legal-content/DE-EN/TXT/?uri=CELEX:32016R0679. (abgerufen am 22.3.2018)

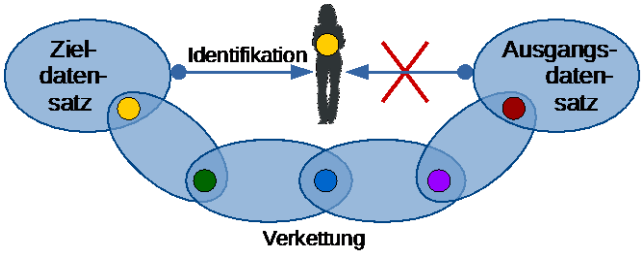

Abbildung 4: Verkettung über mehrere Schritte.

Die Identifikation einer betroffenen Person muss nicht unbedingt mathematisch **eindeutig** sein. Zum Beispiel können biometrische Datensätze Personen nur mit einer gewissen **Wahrscheinlichkeit** identifizieren. Ein ähnliches Konzept sind Mengen von möglichen Betroffenen. Dies ist in Abbildung 5 gezeigt. Insbesondere zeigt die Figur wie ein einzelnes Attribut im Datensatz die Menge aller möglichen Betroffenen einengen kann und damit eine Menge von Kandidaten bestimmt. Durch Verkettung ist es oft möglich, Kandidatenmengen zu überschneiden und damit eine eindeutige Identifikation der betroffenen Person zu erreichen.

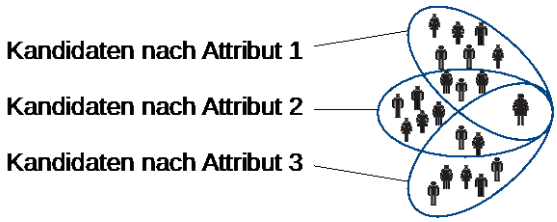

Abbildung 5: Mengen von möglichen Betroffenen.

Das Potential zum Personenbezug eines einzelnen Attributwerts einer Person wächst, je seltener es in der Gruppe von möglichen Personen ist. So hat z.B. die Tatsache, dass eine Person blondes Haar hat ein kleineres Potential zum Personenbezug in Schweden als in Zentral-Afrika.

Aber selbst wenn die möglichen Werte eines Attributs gleichverteilt über die Bevölkerung auftreten, gibt es bestimmte Attribute, die typischerweise für die Identifikation und Verkettung verwendet werden. Diese werden im Folgenden besprochen:

4.1.1 Identifikation über eindeutige Kennungen

Der einfachste Weg zu einem (in einer Domäne) eindeutigen Attribut für eine Person zu kommen, ist die Zuweisung einer eindeutigen Kennung. Diese Praxis war schon vor dem digitalen Zeitalter gängig, ist aber mit Computern und Netzwerken allgegenwärtig geworden.

Ein starker Einflussfaktor der Informatik ist, dass Speicherungsverfahren typischerweise eindeutige Kennungen voraussetzen. So verwenden z.B. Datenbanksysteme sogenannte

Primärschlüssel (primary keys). Ein zweiter Faktor ist die Verfügbarkeit von Netzwerken, dank denen verteilte Stellen die Eindeutigkeit von Kennungen koordinieren können.

Die folgende Tabelle gibt eine Übersicht von einigen gängigen Kennungen. Sie dokumentiert die Vielfalt und Allgegenwärtigkeit solcher Kennungen und illustriert wie eine Person typischerweise eine Vielzahl von Kennungen zugewiesen bekommt.

Ausgeber	***Beispiele***	***Berechtigte***	***Wo eindeutig?***	***Kann ein Inhaber gleichzeitig mehrere haben?***	***Lebensdauer***
Projekt oder System	Primary Key, Benutzername, Pseudonym	Betroffene, Benutzer	Beschränkt auf Projekt oder System	Manchmal ja	Projekt, System
(Private) Organisation	Kunden-Nr. Mitglieds-Nr. Angestellten-Nr. Matrikel-Nr.	Personen mit Beziehung zur Organisation	Beschränkt auf Org.	Je nach Org.	Bis zu Lebenslang
Staat	Steuernr., SSN	Personen: Bürger (Ansässige)	**national**	Nein (Ausnahmen)	Lebenslang
	Ausweis / Pass Nr.		**International** (ICAO)	Limitiert	Auslauf des Dokuments
Techn./org Infrastruktur	E-Mail Adresse pers. Geräte (Tel, MAC, IPv6) Zahlungsmittel	Alle Personen, oft auch Roboter, Hunde, ..	weltweit	Ja	Je nach Ausgeber Lebensdauer, Wechsel von Geräten

Tabelle 1: Überblick von eindeutigen Kennungen.

Vier Charakteristiken sind von besonderem Interesse im Zusammenhang mit Personenbezug. Sie werden im Folgenden diskutiert:

Die Kennung kann nur Personen zugewiesen werden

Entgegen der intuitiven Erwartung können Kennungen, obwohl sie vorrangig von Personen benutzt werden, auch andern Entitäten zugewiesen werden. E-Mail Adressen für Haustiere sind ein Beispiel. Oft ist es aber nicht gewünscht, dass hinter der Kennung kein Mensch steht. Dies ist z.B. der Fall, wenn (Ro)Bots oder Organisationen, oft für illegitime Zwecke, als Personen agieren. Diese Problematik ist gut bekannt und wurde z.B. im folgenden Cartoon[9] dargestellt (siehe Abbildung 6).

Abbildung 6: Cartoon von Peter Steiner publiziert im The New Yorker.

Ein Inhaber (in einer Domäne) **kann nur eine Kennung bekommen**

Oft wird angenommen, dass in einer bestimmten Domäne eine Person nur eine Kennung haben kann. Dies ist vorwiegend der Fall, es gibt aber auch Ausnahmen. Die folgenden Beispiele sollen das illustrieren:

In Italien, erhält man bei Namensänderungen eine neue national eindeutige Steuerkennung. Dies ist darin begründet, dass die Steuerkennung vom Namen abgeleitet wird. Interessanterweise ändern italienische Frauen bei Heirat ihren Namen nicht. Die Änderung der Steuerkennung tritt deshalb am häufigsten bei Heirat von Ausländerinnen auf, die gemäß ihrem anzuwendenden Recht ihre Namen ändern[10].

Weniger häufige Ausnahmen stehen im Zusammenhang mit Zeugenschutzprogrammen oder verdeckten Ermittlungen durch Ordnungshüter. Diese Fälle werden typischerweise auch gesetzlich geregelt.

[9] Quelle: https://en.wikipedia.org/wiki/On_the_Internet,_nobody_knows_you%27re_a_dog
[10] Aussage basiert auf der Arbeitserfahrung des Autors in der Einwohnerkontrolle der italienischen Gemeinde von Grosseto.

Eine Kennung referenziert genau eine Person

Oft wird blind angenommen, dass eine Kennung in einer Domäne immer nur eine einzelne Person referenziert. Auch dazu gibt es Ausnahmen. Beispiele dafür sind die Wiedervergabe von ungenützten Kennungen wie Domänennamen, E-Mail Adressen oder Telefonnummern. Ein weiteres Beispiel ist Geräteweitergabe bei gerätegebundenen Kennungen wie z.B. bei WhatsApp[11].

4.1.2 Identifikation über Namen

Namen sind dafür da, Personen zu adressieren. Adressierbarkeit bedingt auch Eindeutigkeit. Vornamen sind deshalb oft in kleinen sozialen Umfeldern (z.B. in der Familie) eindeutig. Umfelder mit mehreren Personen mit demselben Namen führen oft dazu, dass auch Spitznamen als Rufnamen benützt werden, die dann wieder Eindeutigkeit herstellen.

Während es viele Typen von Familiennamen gibt[12], ist zumindest von einigen anzunehmen, dass sie im sozialen Umfeld bei ihrer Entstehung für eine Familie eindeutig waren. Beispiele dafür sind Berufsnamen[13], wo man annehmen kann, dass in einem Dorf nur eine Familie einen bestimmten Beruf ausübte, und Herkunftsnamen[14], die in einem beschränkten Umkreis typischerweise auch eindeutig sind. Durch die z.T. jahrhundertelange Weitergabe dieser ursprünglichen Familiennamen mit Migration und Umsiedlung kann heute aber keine Eindeutigkeit mehr erwartet werden. Die Eindeutigkeit von Namen kann man über verschiedene Online-Dienste verifizieren. Diese stützen sich typischerweise auf Zensusdaten. Beispiele sind:

- http://howmanyofme.com/
- http://forebears.io/surnames
- http://name-statistics.org/

Auch in Domänen mit hohen Bevölkerungszahlen ist die Kombination von Vor- und Nachname oft schon eindeutig. Beispielsweise gibt es laut HowManyOfMe.com in den U.S.A. 8149 Personen mit Vornamen „Bud“ und 464 Personen mit Nachname „Bruegger“, aber die Kombination „Bud Bruegger“ ist eindeutig. Mit häufigeren Namen ist die Eindeutigkeit aber nicht garantiert. Es gibt z.B. 21 „Donald Trumps“ und 47‘132 „John Smiths“. Trotzdem hat Vor- und Nachname, auch ohne Mittelnamen oder –Initialen, und ohne Alters- oder Ortsangabe schon ein sehr hohes Identifizierungspotential.

[11] https://faq.whatsapp.com/en/general/24460358, Version vom 23.3.2018.
[12] https://en.wikipedia.org/wiki/Surname#Typology, Version vom 27.3.2018.
[13] https://en.wikipedia.org/wiki/Surname#Occupational_surname, Version vom 27.3.2018.
[14] https://en.wikipedia.org/wiki/Surname#Toponymic_surname, Version vom 27.3.2018.

Wie viele eindeutige Namenspaare es in verschiedenen Ländern gibt ist in der folgenden Abbildung 7 dargestellt[15]. Zu beachten ist, dass die Zahlen von eindeutigen Namenspaaren absolut sind. Eine Relativierung mit der Landesbevölkerung wäre vielleicht aussagekräftiger.

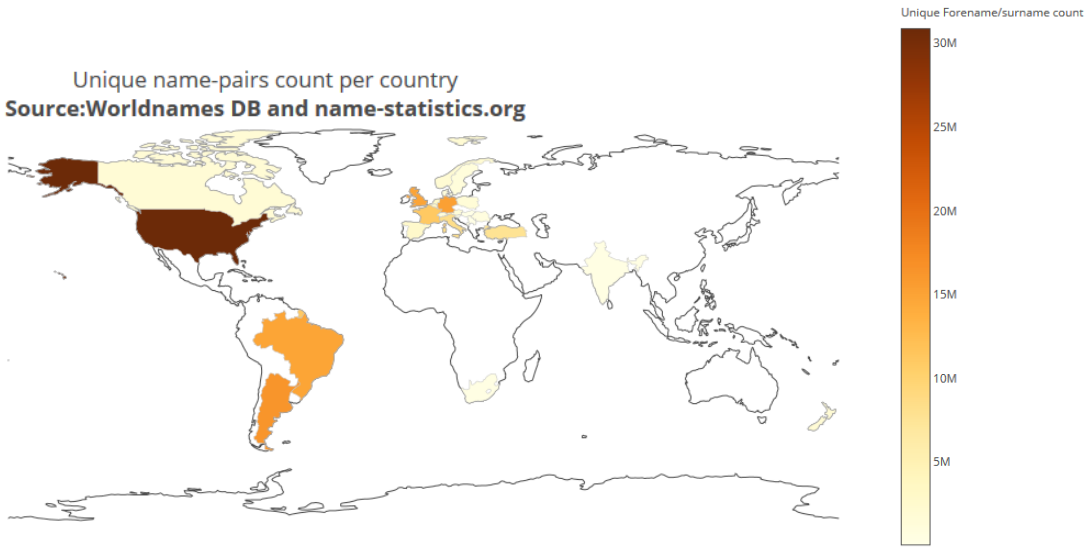

Abbildung 7: Eindeutige Namenspaare es in verschiedenen Ländern gemäß hhainguyen.

4.1.3 Identifikation über Namen, Geburtsort und Datum

Wenn schon Vor- und Nachname ein hohes Identifizierungspotential haben, dann kann die Kombination mit einem Ort und einer Information zum Alter die Wahrscheinlichkeit von Eindeutigkeit sehr hoch treiben.

Dies wird von einigen pre-digitalen nationalen Kennungen ausgenützt. Z.B. wird die alte Schweizer AHV-Nummer vom Beginn des Namens, dem Geburtsdatum und dem Geschlecht abgeleitet[16]. Ähnlich wird der italienische Codice Fiscale aufgrund der (wenn möglich nur) Konsonanten von Vor-und Nachname, dem Geburtsdatum, dem Geburtsort und dem Geschlecht abgeleitet[17].

Beide diese Konzepte haben zusätzlich eine Prüfziffer in die Kennung integriert, um in den sehr seltenen Fällen von Kollisionen trotzdem Eindeutigkeit zu erreichen. Bei dezentral ausgegebenen Kennungen ist aber eine derartige Prüfung der Eindeutigkeit erst seit der Verfügbarkeit von Netzwerken praktisch durchführbar. Vor dieser Möglichkeit scheinen aber die Kennungskonzepte in der Praxis gut funktioniert zu haben.

[15] Quelle: https://plot.ly/~hhainguyen/74/unique-name-pairs-count-per-country-sourceworldnames-db-and-name-statisticsorg/
[16] http://www.ahvnummer.ch/aufbau-alt.htm
[17] https://en.wikipedia.org/wiki/Italian_fiscal_code_card

4.1.4 Identifikation über Postleitzahl, Geschlecht und Geburtsdatum

Vielleicht erstaunlich ist es, dass die Kombination von Postleitzahl, Geschlecht und Geburtsdatum schon hoch identifizierend ist. Isoliert betrachtet werden diese Attribute ja fast als „anonym" eingestuft. Eine Studie[18] von Latanya Sweeney der Carnegie Mellon University zeigt das erstaunlich hohe Identifikationspotential dieser Kombination. Demgemäß können 87% der amerikanischen Bevölkerung eindeutig durch diese drei Attribute identifiziert werden.

Dies zeigt einerseits das große Identifikationspotential einer Orts-Zeit Kombination. Zum andern zeigt es dass die intuitive Einschätzung des Identifikationspotentials nicht immer zutrifft.

4.1.5 Identifikation über Geräte und Dinge

Identifikation einer Person erfolgt oft nicht direkt, sondern über Dinge, die dieser Person zugeordnet werden können. Vor allem Geräte (Devices) sind oft persönlich. Dies wird mit den folgenden Beispielen unterlegt.

Smartphone, Tablet, Telefon, PC, Bloototh Headset

- Telefonnummer, MAC Address, Cookie, Browser-Fingerprint, IP#

E-Mail, Social Media Account (firstname.lastname@domain, aber info@company.com)

- E-Mail Adresse, Benutzername

Fahrzeug (Privatfahrzeug, Dienstwagen, Mietwagen)

- Kennzeichen, Zertifikat in C-ITS, RFID Tag, ..

Gebäude/Wohnung (Wohnort, Arbeitsort, Clubhaus, .., Hotel)

- Postadresse

Bankkonto, Bank- und Kreditkarten

- Nummern

4.1.6 Identifikation über Ort und Zeit

Wie schon oben angedeutet ist die Kombination von Ort und Zeit hoch identifizierend. Dies ist wahrscheinlich nicht überraschend, wenn man in Betracht zieht, dass sich gleichzeitig nur eine Person an einem (genauen) Ort aufhalten kann, da unsere Körper physische Objekte sind, die Raum einnehmen.

Die meisten Menschen haben sehr starke Beziehungen mit einigen Orten. Typischerweise schließt das den (Schlaf-)Wohn- und den Arbeitsort ein.

[18] http://repository.cmu.edu/isr/230/, Version vom 23.3.2018.

Der Ort, wo sich eine Person befindet, gibt sehr oft auch Aufschluss über seine Aktivitäten. Orte sind nicht nur Koordinaten im Raum, sondern sie sind Wohnorte, Geschäfte, Kulturstätten, Krankenhäuser, Sportanlagen, Anwaltskanzleien und sehr vieles mehr. Orte beinhalten deshalb fast immer zusätzliche Informationen über die wieder Verkettet werden kann.

Von mehreren Personen besuchte Orte etablieren auch oft Beziehungen zwischen diesen Personen. Ein Beispiel ist der Aufenthalt in einer Privatwohnung, selbst wenn die Personen sich nicht gleichzeitig dort aufhalten.

Wenn einzelne isolierte Orte schon hoch identifizierend sind, dann sind es Kombinationen von Orten noch viel mehr. Eine Identifizierung ist oft selbst dann möglich, wenn diese Orte von vielen Personen besucht werden.

Noch höher identifizierend sind Zeitreihen von Positionen. Zusätzlich zu der Kombination von Orten beschrieben Zeitreihen auch eine Reihenfolge. Wenn die Auflösung der Zeitreihe hoch genug ist, können sie sogar als Verhaltensmuster angesehen werden, die sehr persönlich sein können. Ein Beispiel dafür sind Zeitreihen einer Autofahrt. Die daraus ableitbaren Beschleunigungsmuster geben Auskunft über den Fahrstil anhand dessen man Personen identifizieren kann.

Von zwei Personen geteilte Zeitreihen etablieren typischerweise sehr starke Beziehungen. Dies selbst an öffentlichen Orten, die von vielen Personen besucht sind.

4.1.7 Identifikation über von Ort und Zeit abgeleitete Daten

Da Ort und Zeit so stark identifizierend sind, wird oft versucht, solche Daten zu anonymisieren. Dabei können z.B. die Genauigkeit des Ortes vergröbert werden oder Zeitreihen zu Beschleunigungsmustern vereinfacht werden. Studien zeigen aber, dass entgegen unserer Intuition solche „Anonymisierungsmethoden“ eine Identifikation der Person nicht verhindern können.

In einem Beispiel wurde eine Orts-Zeitreihe so reduziert, dass nur noch die Mobilzelle jede Stunde bekannt war. In ihrem Artikel in Nature schreiben Montjoye, Hidalgo, Verleysen und Blondel: „*In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals.*“[19]

[19] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel, Unique in the Crowd: The privacy bounds of human mobility, Nature, Scientific Reports 3, Article number: 1376 (2013), http://rdcu.be/JGxh

Unter dem aufschlussreichen Titel „**Anonymization of location data does not work: A large-scale measurement study**"[20] schreiben Zang und Bolot:

> "*We examine a very large-scale data set of more than* ***30 billion call records made by 25 million cell phone users across all 50 states*** *of the US and attempt to determine to what extent anonymized location data can reveal private user information. Our approach is to* ***infer****, from the call records, the* ***"top N" locations for each user*** *and correlate this information with publicly-available side information such as census data. For example, the measured "top 2" locations likely correspond to home and work locations, the "top 3" to home, work, and shopping/school/commute path locations. We consider the cases where those* ***"top N" locations are measured with different levels of granularity, ranging from a cell sector to whole cell, zip code, city, county and state. We then compute*** *the anonymity set, namely the* ***number of users uniquely identified by a given set of "top N" locations at different granularity levels. We find that the "top 1" location does not typically yield small anonymity sets. However, the top 2 and top 3 locations do, certainly at the sector or cell-level granularity.*** *We consider a variety of different factors that might impact the size of the anonymity set, for example the distance between the "top N" locations or the geographic environment (rural vs urban). We also examine to what extent specific side information, in particular the size of the user's social network, decrease the anonymity set and therefore increase risks to privacy.* ***Our study shows that sharing anonymized location data will likely lead to privacy risks*** *and that, at a minimum, the data needs to be coarse in either the time domain (meaning the data is collected over short periods of time, in which case inferring the top N locations reliably is difficult) or the space domain (meaning the data granularity is strictly higher than the cell level). In both cases, the utility of the anonymized location data will be decreased, potentially by a significant amount.* " (Hervorhebung durch den Autor)

Unter dem Titel "**Elastic Pathing: Your Speed is Enough to Track You**"[21] schreiben Gao, Firner, Sugrim, Kaiser-Pendergrast, Yang und Lindqvist:

> "*Today, people have the opportunity to opt-in to usage-based automotive insurances for reduced premiums by allowing companies to monitor their driving behavior. Several companies claim to* ***measure only speed data to preserve privacy****. With our elastic pathing algorithm, we show that drivers can be tracked by merely collecting their speed data and* ***knowing their home location, which insurance companies do****, with an accuracy that constitutes privacy intrusion. To demonstrate the algorithm's real-world applicability, we evaluated its performance with datasets from central New Jersey and Seattle, Washington, representing suburban and urban areas.* ***Our algorithm predicted destinations with error within 250 meters for 14% traces and within 500 meters for 24% traces*** *in the New Jersey dataset (254 traces). For the Seattle dataset (691 traces), we similarly predicted destinations with error within 250 and 500 meters for 13% and 26% of the traces respectively.* ***Our work shows that these in-***

[20] Zang, Bolot: Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study, MobiCom 2011, https://www.researchgate.net/publication/220926571_Anonymization_of_location_data_does_not_work_A_large-scale_measurement_study

[21] Gao et al.: Elastic Pathing: Your Speed is Enough to Track You, UbiComp 2014, https://www.winlab.rutgers.edu/~janne/elasticpathing-ubicomp14.pdf

> ***surance schemes enable a substantial breach of privacy***." (Hervorhebung durch den Autor)

Unter dem Titel "**Inferring Trip Destinations From Driving Habits Data**"[22] schreiben Dewri, Annadata, Eltarjaman und Thurimella:

> „The collection of driving habits data is gaining momentum as vehicle telematics based solutions become popular in consumer markets such as auto-insurance and driver assistance services. These solutions rely on **driving features such as time of travel, speed, and braking** to assess accident risk and driver safety. **Given the privacy issues surrounding the geographic tracking of individuals**, many solutions explicitly claim that the customer's **GPS coordinates are not recorded**. Although revealing driving habits can give us access to a number of innovative products, we believe that the disclosure of this data **only offers a false sense of privacy**. **Using speed and time data** from real world driving trips, we show that the **destinations of trips may also be determined** without having to record GPS coordinates. Based on this, we argue that customer privacy expectations in non-tracking telematics applications need to be reset, and new policies need to be implemented to inform customers of possible risks." (Hervorhebung durch den Autor)

4.1.8 Identifikation über Biometrie

Biometrische Merkmale sind eine weitere Möglichkeit Personen zu identifizieren. Dazu werden verschiedene Merkmale einzeln oder kombiniert (sogenannt „multimodal") verwendet.

Diese Merkmale sind entweder Eigenschaften des Körpers wie:

- Gesicht
- Fingerabdruck
- Irismuster
- Ohrgeometrie
- Handgeometrie
- Fingervenen
- Stimme

oder Merkmale, die das Verhalten einer Person beschreiben (sogenannte dynamische Biometrie):

- Herzschlag

[22] Dewri et al.: Inferring Trip Destinations from Driving Habits Data, WPES 2013, http://cs.du.edu/~rdewri/data/MyPapers/Conferences/2013WPES-Extended.pdf

- Tippverhalten (Timing)
- Handschrift (Druck)
- Gang (Beschleunigung)
- Auto Fahrstil (Beschleunigung)
- Ausdrucksweise/
- Wortwahl

Identifikation von Personen durch Biometrie ist immer mit einer Wahrscheinlichkeit behaftet. Dies ist in Abbildung 8 illustriert[23]: Sie zeigt wie die gemessenen Werte eines biometrischen Merkmals bei mehrfachen Messungen schwanken und mit einer gewissen Wahrscheinlichkeit eintreffen. Die grüne Glocke zeigt die Wahrscheinlichkeitsverteilung des legitimen Inhabers der Biometrie, die rote diejenige einer anderen Person. Da diese zwei Verteilungen (Glocken) überlappen, kann ein vorliegender Messwert beiden Personen zugewiesen werden. Wenn man fälschlich schließt, den legitimen Inhaber vor sich zu haben, spricht man von einem „false accept"; wenn fälschlich geschlossen wird, dass es nicht der legitime Inhaber ist, spricht man von einem „false reject".

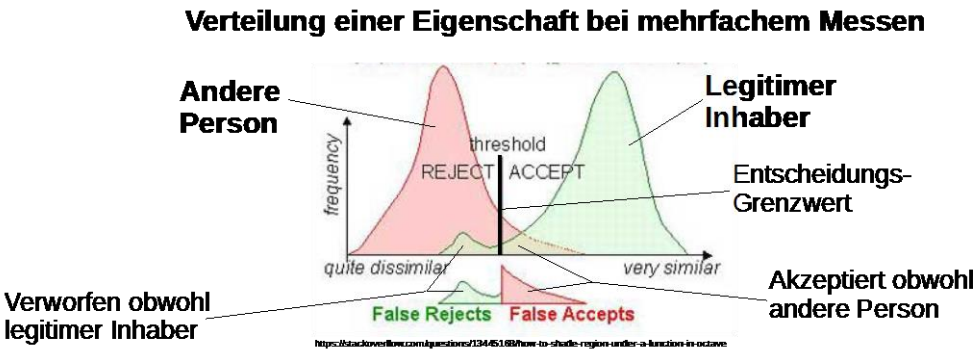

Abbildung 8: Wahrscheinlichkeit der Identifizierung der richtigen Person aufgrund von biometrischen Eigenschaften.

4.2 Verkettungspotential

Neben der Stärke des Bezugs zu der betroffenen Person beeinflusst auch die Stärke des Bezugs zu anderen Datensätzen das Risiko. Dies ist z.B. ersichtlich, wenn ein Datensatz eine Identifizierung der betroffenen Person nicht ermöglicht und diese nur über die Verkettung mit andern Datensätzen möglich wird. Es ist offensichtlich, dass das Risiko sinkt, je schwieriger eine Verkettung möglich ist.

[23] Quelle der Graphik: Francisco Presencia, https://stackoverflow.com/questions/13445168/how-to-shade-region-under-a-function-in-octave

Von diesem Gesichtspunkt stellen Datensätze mit eindeutigen Kennungen ein hohes Risiko für die Rechte und Freiheiten der betroffenen Personen dar. Das Risiko kann noch weiter steigen, wenn ein Datensatz mehrere eindeutige Kennungen enthält. In diesem Falle ist der Datensatz, unabhängig von anderem Inhalt, attraktiv für Angreifer, die ansonsten separate Datensätze verknüpfen wollen um einen breiteren Einblick ins Leben der Betroffenen gewinnen wollen.

In EIDI sollen deshalb Datensätze mit mehreren hoch-identifizierenden Merkmalen besonders geschützt werden. Wiederum ist Pseudonymisierung eine der möglichen Schutzmaßnahmen. Damit wird verhindert, dass EIDI selbst zu einem attraktiven Angriffspunkt wird, der zumindest kurzzeitig das Risiko für Betroffene über das der Leaks hinaus weiter erhöht.

Im Folgenden werden Beispiele für Datensätze gegeben, die mehrere Kennungen zusammenführen:

Das erste Beispiel betrifft Verkettungsdaten, die aus durchaus legitimen Sicherheitsgründen gesammelt werden. Insbesondere basiert es auf einer Maßnahme zur Bekämpfung des Identitätsdiebstahls bei Online Diensten. Sie wird „Risiko-basierte Authentisierung“ genannt. Dazu wird präventiv gespeichert, mit welchen Endgeräten und von welchen Orten die Anmeldung an den Dienst erfolgt. Die Abbildung 9 zeigt, wie diese Daten einsehbar sind[24]. Mit großer Wahrscheinlichkeit werden auch identifizierende Charakteristiken, wie Cookies und Fingerprints, dieser Endgeräte gespeichert.

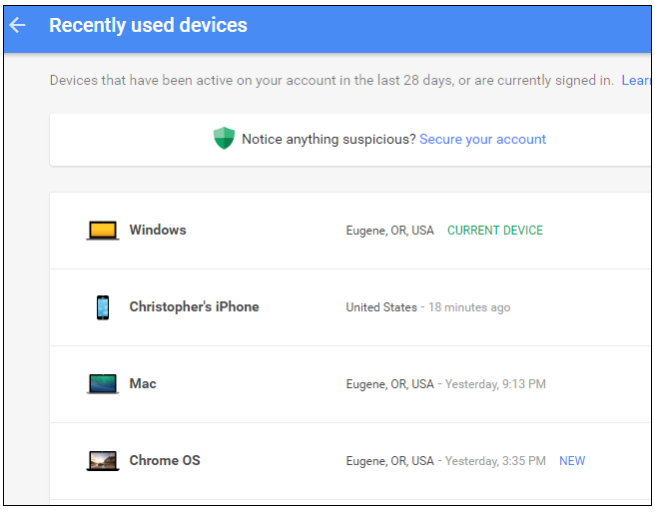

Abbildung 9: Verkettung von Benutzerendgeräten.

Wenn der Benutzer sich dann mit einem anderen Endgerät oder von einem ungewohnten Ort anmeldet, wird das als erhöhtes Risiko gesehen, dass es sich um einen Identitätsdieb

[24] Quelle: https://www.howtogeek.com/279518/how-to-see-other-devices-logged-in-to-your-google-account/

handelt. Um dieses Risiko zu minimieren, wird der Benutzer auf einem anderen Kanal gefragt, ob die Anmeldung wirklich legitim sei. (Siehe Abbildung 10[25]).

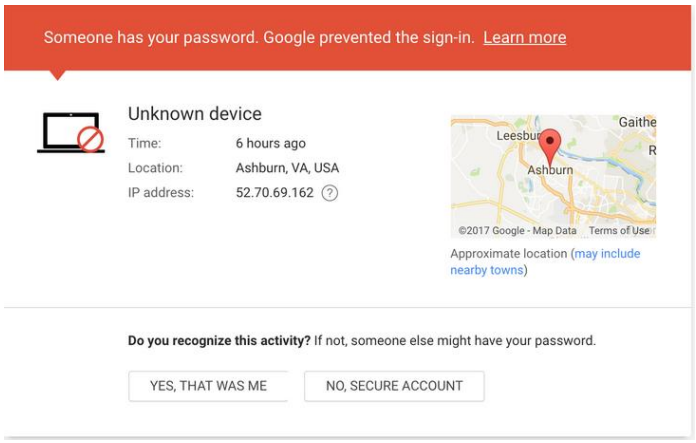

Abbildung 10: Dialog zur Erfassung eines zusätzlichen Benutzergeräts.

Mit dieser Methode bestätigt der Betroffene die Verkettung seiner verschiedenen Kennungen, insbesondre E-Mail Adressen und Telefonnummern. Dies kann natürlich weiter mit den Kennungen von Endgeräten, die auf den Benutzer-Account zugreifen, verkettet werden.

Ein weiteres Beispiel sind sogenannte „Location Services" die typischerweise von Smartphone-Betriebssystemen genutzt werden, um die eigene Position genau feststellen zu können. Während GPS direkt Positionskoordinaten (Lat/Long) ausgibt, ist guter GPS Empfang nicht immer garantiert. In Empfangslöchern von GPS Signalen, z.B. im Innern von Gebäuden, muss deshalb eine andere Art der Positionierung benutzt werden. Diese basiert typischerweise auf der Signalstärke von WiFi und/oder Mobilfunknetzen. Diese ermöglicht eine Triangulation zwischen den Positionen der Wifi-Zugangspunkte oder Masten der Mobilfunkzellen. Diese Positionen sind aber nur selten öffentlich erhältlich.

Aus diesem Grund bieten Firmen, wie z.B. Google[26], Location Services an. Dies basiert auf einer Datenbank von Positionen von Wifi-Zugangspunkten und Mobilfunkmasten, die durch Crowdsourcing erfasst wurden, indem teilnehmende Smartphones sowohl GPS Koordinaten, als auch Signalstärken von sichtbaren Netzen mit dem Betreiber des Location Services geteilt haben.

Bei der Benutzung solcher Location Services gibt der Betroffene zwangsweise sein Bewegungsprofil preis. Kombiniert mit dem ersten Beispiel ist also ersichtlich, wie stark identifizierend auch eine einzige Position des Betroffenen sein kann.

---

[25] Quelle: https://docs.bitnami.com/virtual-machine/apps/roundcube/

[26] https://developers.google.com/android/reference/com/google/android/gms/location/LocationServices

Ein drittes Beispiel ist das Tagging[27] von Personen auf (Facebook) Fotos. Hier werden durch Crowdsourcing typischerweise Gesichter in Fotos mit der eindeutigen Facebook-Kennung verknüpft. Dies erzeugt nicht nur Daten über die Beziehung des Taggers mit der getaggten Person, sondern erstellt auch einen reichen Datensatz zur Verkettung von Gesichtsbiometrie mit Personenkennungen. Diese Daten können dann natürlich auch für das Training von automatischen Gesichtserkennungsalgorithmen[28][29] eingesetzt werden.

Diese Beispiele zeigen, dass die Verfügbarkeit von Verkettungsdatensätzen durchaus bei der Risikoabschätzung in Betracht gezogen werden muss. Nicht erstaunlich verketten diese Datensätze die stark identifizierenden Attribute wie verschiedene Kennungen, verschiedene Geräte, Gesichtsbiometrie und Ort/Zeit.

4.3 Potential der Zweckentfremdung

Gemäß DS-GVO ist die Verarbeitung von personenbezogenen Daten nur dann erlaubt, wenn die Daten für einen erklärten und dem Betroffenen bekannten Zweck erhoben und verarbeitet werden[30]. Somit ist auch ein möglicherweise eingeholtes Einverständnis des Betroffenen exklusiv auf diesen Zweck beschränkt. Die Verwendung der erhobenen Daten für andere Zwecke und durch nicht vom Betroffenen autorisierte Parteien kann einen erheblichen Eingriff in die Rechte und Freiheiten der betroffenen Person darstellen. Insbesondere verlieren Betroffene damit jegliche Kontrolle darüber, wer ihre Daten für welche Zwecke verwendet.

Zweckentfremdungen kreieren das Risiko, dass der Betroffene materiellen oder immateriellen Schaden erleidet. Ein Beispiel dafür ist die Verwendung von Daten, die ursprünglich in einem begrenzten Kreis von Freunden vom Betroffenen geteilt wurden, von einem potentiellen Arbeitsgeber zur Auswahl von Kandidaten oder von einem Kreditinstitut für die Bestimmung der Kreditwürdigkeit.

Das Potential zur Zweckentfremdung von Daten wird davon bestimmt, wie wertvoll diese Daten für andere Zwecke sind. Dies soll mit den folgenden zwei Beispielen illustriert werden: Die Schuhgröße eines Betroffenen ist wahrscheinlich nur in einem relativ engen Segment von Zwecken relevant. Die Postleitzahl hingegen erlaubt Rückschlüsse, die für eine

[27] https://www.facebook.com/about/tagging
[28] Sidney Fussell, Facebook's New Face Recognition Features: What We Do (and Don't) Know [Updated], 27.2.2018, https://gizmodo.com/facebooks-new-face-recognition-features-what-we-do-an-1823359911, in der Version vom 26.3.2018.
[29] Joaquin Quiñonero Candela, Managing Your Identity on Facebook with Face Recognition Technology, December 19, 2017, https://newsroom.fb.com/news/2017/12/managing-your-identity-on-facebook-with-face-recognition-technology/
[30] Siehe DS-GVO Art. 5(1)(b).

Vielzahl von Zwecken nützlich sein kann. Die Information über den Wohnort kann den Betroffenen z.B. als möglichen Interessenten für eine Vielzahl von ortsgebundenen Angeboten für ein weites Spektrum von Dienstleistungen identifizieren. Darüber hinaus könnte es Rückschlüsse über Einkommen, Bildung oder wahrscheinlichen Familienstand geben. Dies kann dann z.B. auch für die Bestimmung von Versicherungsprämien oder Kreditwürdigkeit eingesetzt werden.

Das Potential zur Zweckentfremdung von Datenelementen kann nur im konkreten Fall eingeschätzt werden. Die Einschätzung dieses Risikos muss deshalb separat für jedes von EIDI verarbeitetes Datenelement getätigt werden.

Die Identifikation und der Schutz der Datenelemente mit hohem Potential für Zweckentfremdung dienen dazu, dass EIDI nicht selbst als Angriffspunkt für Angreifer attraktiv oder selbst zum Angreifer wird.

4.4 Risikopotential durch die Art der Daten

Verschiedene Arten von Daten informieren über verschiedene Aspekte der Betroffenen. Gewisse Aspekte beinhalten ein besonders hohes Risikopotential, entweder weil sie über sehr intime Bereiche des Lebens Einsicht gewähren oder anderweitig ein hohes Potential für Missbrauch wie z.B. Diskriminierung tragen.

Artikel 9(1) der DS-GVO[31] beschreibt Merkmale, die besonders schutzwürdig sind. Diese sind Daten „*aus denen die rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen",* sowie „*genetischen Daten, biometrischen Daten zur eindeutigen Identifizierung einer natürlichen Person, Gesundheitsdaten oder Daten zum Sexualleben oder der sexuellen Orientierung einer natürlichen Person*". Artikel 10 führt zusätzlich „*Daten über strafrechtliche Verurteilungen und Straftaten oder damit zusammenhängende Sicherungsmaßregeln*" auf.

Solche Daten sind durchaus in verfügbaren Daten-Leaks enthalten. Das beträchtliche Risiko dieser Daten für die Rechte und Freiheiten der Betroffenen stammt vorwiegend aus der Vertraulichkeit, die durch den Breach beim ursprünglichen Verantwortlichen verloren gegangen ist.

Dieses Risiko für die Betroffenen könnte lediglich durch eine Löschung der Daten durch die Betreiber der Datensenken reduziert werden. EIDI kann deshalb allenfalls eine entsprechende Benachrichtigung dieser Betreiber in Betracht ziehen. Dazu ist es lediglich notwendig zu wissen, dass ein Datensatz solche Merkmale enthält. Die eigentlichen Werte dieser Merkmale sind für diesen Zweck völlig unnötig. Intime Daten sind auch unnötig für eine mögliche Warnung der Betroffenen.

[31] https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:32016R0679&from=DE

Aus diesen Gründen soll EIDI möglichst vermeiden, diese Art von Daten zu erheben, zu speichern oder anderweitig zu verarbeiten.

4.5 Potential von Folgeschäden

Bisher wurde das Risiko betrachtet, dass durch die direkte Verwendung von schon geleakten Daten entstanden ist. Ein zusätzliches Risiko besteht, wenn diese Daten Zugang zu derzeit noch geschützten Daten und Werten ermöglichen. Dies ist insbesondere für die folgenden Daten der Fall:

- Daten, die für eine Vortäuschung einer anderen Identität (englisch: „impersonation“) dienen können,
- Daten, die Zugang zu Zahlungsmitteln gewähren,
- Daten, die Zugang zu Online-Konten ermöglichen.

Im ersten Fall handelt es sich um einen Datensatz der dafür genügt, glaubhaft eine andere Identität anzunehmen. Dazu sind typischerweise Daten wie Name, Adresse und Geburtsdatum notwendig. In den Vereinigten Staaten wird für diese Fälle auch typischerweise die Social Security Number und der Mädchenname der Mutter genannt. Diese Art von Daten können Zugang zu sehr schwach geschützten Daten und Werten gewähren.

Da es für EIDI wahrscheinlich unmöglich ist, die Verfügbarkeit dieser Daten in Datensenken zu verhindern, muss sich das Projekt darauf beschränken, mit geeigneten Maßnahmen zu verhindern, dass diese Daten auch von EIDI gestohlen und weiter verbreitet werden können.

Interessanter für EIDI sind die letzteren zwei Fälle, bei denen Leaks eigentliche Kredentiale enthalten. Dabei handelt es sich typischerweise um Daten von Kreditkarten[32] und Benutzername/Passwort für den Zugang zu Online-Konten. In beiden Fällen kann EIDI effektive Maßnahmen zur Begrenzung des Risikos der Betroffenen implementieren.

Insbesondere kann EIDI bei Kreditkarteninformation den Betroffenen oder den Ausgeber der Karte warnen und damit eine Sperrung der Karte veranlassen. Bei geleakten Passwörtern kann EIDI sowohl den Betroffenen warnen, das Passwort zu ändern, als auch den Betreiber des Online-Dienstes informieren, so dass das Konto ohne zusätzliche Legitimation gesperrt wird.

[32] Insbesondere sind Kreditkartennummer, Verfallsdatum, Name des Halters und die Prüfziffer ausreichend um Zahlungen zu tätigen.

4.6 Verfügbarkeit und Auffindbarkeit der Daten

Die Verfügbarkeit hat einen direkten Einfluss auf die Schwere des Risikos durch einen Daten-Breach. Je weniger Personen Zugriff zu den Daten haben, desto geringer das Risiko eines Missbrauchs und Schadens.

Die Verfügbarkeit wird von den folgenden Faktoren beeinflusst:

- Anzahl der Datensenken über die die Daten verfügbar sind,
- die Anzahl und Motivation der Personen, die auf die Datensenken Zugriff haben,
- das Zeitfenster in dem die Daten zugreifbar sind und
- die Leichtigkeit, mit denen die Daten auffindbar sind[33].

Offensichtlich steigt mit der Anzahl der Datensenken die Wahrscheinlichkeit, dass die Daten gefunden werden. Mit steigender Anzahl von Datensenken steigt die wahrscheinliche Anzahl von Personen, die auf die Daten zugreifen.

Je länger Daten auf einer gegebenen Datensenke verfügbar sind, desto größer ist die Zahl der Personen, die von der Verfügbarkeit erfahren (z.B. durch Weiterempfehlung) und desto größer die Zahl der heruntergeladenen Kopien.

5. Risiken, die nicht mehr durch EIDI beeinflusst werden können.

Das Ziel von EIDI ist die Verringerung der Risiken für die Rechte und Freiheiten von betroffenen Personen nach einem Breach. Einige durch einen Breach entstandene Risiken sind irreversibel, können also nicht mehr durch EIDI beeinflusst werden. Dies sind insbesondere Risiken die durch die Verletzung der Vertraulichkeit entstanden sind.

Um die verlorene Vertraulichkeit wiederherzustellen, müsste es möglich sein, alle illegitimen Kopien eines Leaks zu löschen. Aber selbst wenn das für alle möglichen Datensenken gleich nach dem Auftauchen eines Leaks möglich wäre, dann könnte man trotzdem nicht vermeiden, dass die ursprünglichen Datendiebe und –Hehler Kopien der Daten behalten. Außerdem können selbst bei schnellem Einschreiten neue Kopien heruntergeladen werden. Die

[33] So sind z.B. einige Leaks über Suchmaschinen auffindbar, andere nur durch Mundpropaganda.

Erfahrung zeigt, dass dieselben Daten regelmäßig erneut in Datensenken auftauchen (siehe D2.1).

Eine vollständige Wiederherstellung der Vertraulichkeit nach einem Breach ist deshalb in vielen Fällen nicht mehr möglich.

6. Risiken, die von EIDI beeinflusst werden können

Angesichts der Tatsache, dass immer wieder Unbefugte Zugang zu den geleakten Daten haben werden, stellt sich natürlich die Frage, wie EIDI trotzdem die Risiken der Betroffenen reduzieren kann. Hier stehen einige Möglichkeiten offen:

6.1 Verringerung der Verfügbarkeit

Sektion 4.6 beschreibt, wie die Verfügbarkeit von geleakten Daten direkt das Risiko beeinflusst. EIDI kann damit das Risiko senken, indem es die Verfügbarkeit reduziert.

Dies kann normalerweise nur durch eine Zusammenarbeit mit den Betreibern von Datensenken gelingen. Dabei kann EIDI Daten-Leaks in der Datensenke als solche identifizieren und den Betreiber benachrichtigen. Der Betreiber der Datensenke kann dann die illegitim gehosteten Daten löschen. Es wäre denkbar, dass der Betreiber die Personen kennt, die die Daten heruntergeladen haben. In diesem Falle könnten auch sie aufgefordert werden, die Daten zu löschen.

Daten-Leaks von digitalen Identitäten, also personenbezogenen Daten, können nicht legal von Betreibern der Datensenken gespeichert werden. Sobald einer der vom Leak betroffenen Personen oder der Verantwortliche in Europa ansässig ist, ist die DSGVO anwendbar. Die Speicherung der Daten in einer -Datenquelle kann keine legitime Rechtsgrundlage gemäß Artikel 6 der DSGVO haben.

Da die Betreiber von Datenquellen durch gespeicherte Leaks das Risiko eingehen, mit einem Bußgeld belegt zu werden, könnte es durchaus in ihrem Interesse sein, mit EIDI zusammenzuarbeiten.

6.2 Strafverfolgung von Identitätsdieben und –Hehlern

Eine mögliche Abschreckungsmaßnahme von EIDI könnte die Wahrscheinlichkeit von Datendiebstahl, -Hehlerei und die Veröffentlichung von Leaks in Datensenken senken. Dies ist möglich, wenn das Hochladen von Leak-Daten auf Datensenken nicht anonym erfolgen kann. In diesem Fall kann EIDI dazu beitragen, die Identität des Datendiebes oder –Hehlers zu ermitteln und damit die zuständigen Strafverfolgungsbehörden zu unterstützen.

6.3 Informierte Schadensbegrenzung durch den Betroffenen

Ein Daten-Breach verletzt auch das Anrecht auf Transparenz des Betroffenen. Personenbezogene Daten können danach von unbekannten Parteien für nicht deklarierte Zwecke benutzt werden. Die Warnung von Betroffenen kann zwar die Transparenz verbessern, aber kann meist nicht aufdecken, wer unrechtmäßig die geleakten Daten zu welchen Zwecken verwendet. Eine signifikante Verringerung des Risikos für Betroffene entsteht erst, wenn die Information zur Schadensbegrenzung genutzt werden kann.

Insbesondere können Betroffene dann oft Schadensbegrenzungsmaßnahmen ergreifen, wenn sie von einem Breach und den damit verbundenen Risiken informiert werden. Diese Strategie vermeidet vor allem, dass der Betroffene unvorbereitet und überraschend zu Schaden kommen kann. Während EIDI oft die ursprüngliche Verarbeitung nicht identifizieren kann, und z.T. über die Semantik von Datenfeldern im Dunkeln bleibt, kennt der Betroffene meist den Informationsinhalt und Kontext eines Breaches genau. Dies ermöglicht es dem Betroffenen, gezielte und persönlich angepasste Schadensbegrenzungsstrategien zu entwickeln. Dies wird von EIDI dadurch ermöglicht, dass die Betroffenen über den Breach wissen und das daraus folgende Risiko verstehen.

6.4 Vermeidung von Folgeschäden

Im Fall der Vermeidung von Folgeschäden existieren Daten oder Werte, deren Vertraulichkeit oder Unversehrtheit noch in Takt ist, wo aber die geleakten Daten Zugang verschaffen können.

Dieses Risiko nur kann dann verkleinert werden, wenn es möglich ist, den Zugang zu sperren und die Zugangsdaten zu ändern. Dies ist z.B. nicht der Fall, wenn Trickbetrüger einen genügenden Umfang von Daten einer Person kennen, um glaubhaft eine falsche Identität anzunehmen. Die für diese Art von „Zugang“ typischerweise benutzten Daten, wie Namen,

Geburtsdatum und –Ort, genaue Adresse und vielleicht nationale Kennungen[34] können aber üblicherweise nicht geändert werden.

EIDI muss sich deshalb auf Zugangsdaten konzentrieren, die geändert werden können. Typische Beispiele dafür sind Passwörter und Kreditkarten. Diese können sowohl durch den Betroffenen als auch durch den Betreiber des entsprechenden Dienstes oder Ausgeber geändert werden.

Bei Passwörtern ist in Betracht zu ziehen, dass viele Benutzer ein Passwort bei mehreren Diensten anwenden. Es können dadurch Risiken bei Diensten entstehen, die nicht vom Breach betroffen sind.

Risikosenkende Maßnahmen durch EIDI müssen weiter auch die Möglichkeit betrachten, dass die Zugangsdaten schon von einem Angreifer kompromittiert wurden und dass der Angreifer jetzt das Konto beim entsprechenden Dienst (oder den E-Mail Account des legitimen Benutzers) kontrolliert. Hier sind Maßnahmen, die sich an den Dienst statt dem Betroffenen richten oft besser geeignet. Der Dienst hat dann Möglichkeiten, die Legitimität des Benutzers zu verifizieren, die EIDI nicht zur Verfügung stehen.

7. EIDI Maßnahmen und Verkettung von Datensätzen

Die Teilvorhabensbeschreibung stellt die Frage[35], „*ob bereits auf Basis der vorliegenden Informationen aus der Datensenke eine Unterrichtung möglich ist oder ob eine Verkettung mit andern Datenbeständen erforderlich ist*“. Diese Frage wird im Folgenden näher betrachtet. Die Frage wird im Sinne der möglichen EIDI Maßnahmen 6.3 und 6.4 oben betrachtet. Dabei wird entweder die betroffene Person (6.3 und 6.4) oder aber der Betreiber des Dienstes, wo Folgeschäden entstehen können (6.4) kontaktiert. Die Frage wird hier deshalb in einem weiteren Sinn diskutiert, wann ein geeigneter Kanal entweder zu der betroffenen Person oder dem betroffenen Dienst verfügbar ist.

Wann ein Kanal geeignet ist, ist nicht leicht zu beantworten. Diese Frage wird in andern Teilarbeitspaketen beantwortet und von der Diskussion hier weitgehend ausgeschlossen.

[34] Anmerkung: Nationale Kennungen zur Identifizierung sind in Deutschland weniger üblich als in andern Ländern. In den Vereinigten Staaten wir z.B. die Social Security Number ein wichtiger Bestandteil von vielen betrügerischen Tätigkeiten.
[35] Sektion 3.1.2.2., Seiten 13 und 14.

Wichtig für das Verständnis ist aber, dass der Sender der Benachrichtigung sehr wichtig für die Effizienz der Maßnahme ist. Deshalb werden auch Szenarien in Betracht gezogen, wo nicht der EIDI Identitätsanalyst selbst als Sender agiert.

Im Folgenden werden die verschiedenen möglichen Szenarien mit und ohne Verkettung dargestellt.

Abbildung 11 zeigt den Fall, wo eine geleakte digitale Identität schon Daten zur Verwendung eines geeigneten Kanals beinhaltet und vom EIDI-Identitätsanalyst als effektiv bewertet wurde. Dann kann der Identitätsanalyst direkt die betroffene Person benachrichtigen.

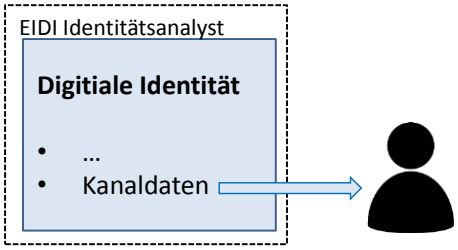

Abbildung 11: Direkte Warnung ohne Verkettung.

Abbildung 12 zeigt den Fall, wo eine erste digitale Identität keine Daten zum Zugang zu einem geeigneten Kanal beinhaltet, diese aber in einer zweiten digitalen Identität desselben Betroffenen zur Verfügung stehen. Die zwei digitalen (Teil-)Identitäten stammen dann von zwei verschiedenen Leaks. In diesem Fall müssen die beiden digitalen (Teil-)Identitäten verkettet werden, wenn der Betroffene über die Risiken durch die erste Identität gewarnt werden soll.

Eine derartige Verkettung basiert typischerweise auf der Verfügbarkeit derselben Kennung in beiden Teilidentitäten. Aber auch wenn die Kennung in den beiden Teilidentitäten identisch ist, besteht immer noch ein gewisses Risiko, dass sich die Teilidentitäten auf verschiedene Personen beziehen. Dies wurde in Sektion 4.1.1. oben schon angesprochen. Hier sollen noch einmal die Beispiele der erneuten Ausgabe einer E-Mail-Adresse nach längerem Nichtgebrauch und die Weitergabe eines persönlichen Geräts und damit dessen Kennung in Erinnerung gerufen werden.

Um das Risiko zu reduzieren, Teilidentitäten von verschiedenen Personen zu verketten, können weitere Daten in den beiden Identitäten auf Gleichheit getestet werden. In der Figur werden diese Daten als „Bestätigungsdaten“ bezeichnet. Beispiele umfassen Vornamen und Geschlecht. Diese sind offensichtlich nicht als Kennung geeignet, können aber die Verkettung von Teilidentitäten verschiedener Personen durchaus aufdecken.

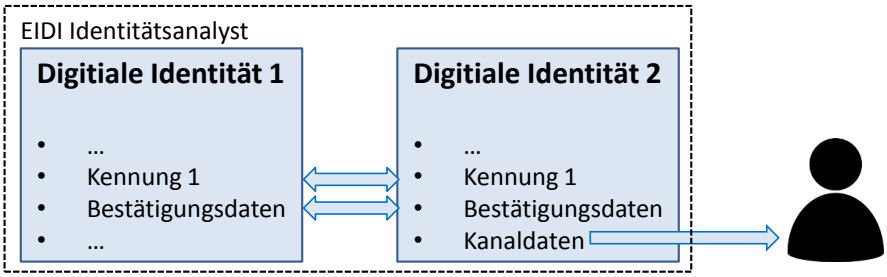

Abbildung 12: Direkte Warnung mit Verkettung.

Abbildung 13 illustriert den Fall, wo der für geeignet befundene Kanal durch einen externen Dienst kontrolliert wird. Ein typisches Beispiel wäre der Kanal einer Login-Nachricht, also eine Meldung, die beim Anmelden des Betroffenen beim externen Dienst angezeigt wird. In diesem Fall ist eine Verkettung der Teilidentität beim EIDI Identitätsanalysten mit dem Benutzerkonto desselben Betroffenen notwendig. Wie im vorhergehenden Fall basiert dies wiederum auf identischen Werten einer Kennung und optionalen Bestätigungsdaten.

Während Abbildung 13 sehr ähnlich zu Abbildung 12 scheint, zeigt sie doch eine ganz andere Situation. Insbesondere hat in Abbildung 12 eine einzige Partei, namentlich der EIDI Identitätsanalyst, Kenntnis und Kontrolle über beide Teilidentitäten und kann die Verkettung durch einfachen Vergleich von Kennungen herstellen. Im Gegensatz dazu hat in Abbildung 13 jede Partei nur Kenntnis über die von ihnen verwalteten Teilidentitäten. Ein freier Austausch von Daten zwischen diesen Parteien würde jeglichem Datenschutz- und Sicherheitsprinzip widersprechen. Die Verkettung kann also nicht durch einfachen Vergleich von Kennungen erreicht werden.

Stattdessen unterliegt die Verkettung einer strengen Anforderung: Keiner der beiden Parteien soll zusätzliche Kenntnis über den Betroffenen erlangen, wenn dies nicht für die Maßnahme notwendig ist, die das Risiko des Betroffenen senkt.

Diese Anforderung hat mehrere Konsequenzen, inklusive der Folgenden:

- Der Warndienst soll keine Information über Personen erhalten, deren Teilidentität zwar in einem Leak enthalten ist, aber die kein Benutzerkonto beim Warndienst haben.
- Der Identitätsanalyst soll nicht erfahren, welche Personen Benutzerkonten beim Warndienst haben.
- Der Warndienst soll keine zusätzlichen Informationen über einen zu warnenden Betroffenen erhalten, außer wenn dies zur Benachrichtigung direkt notwendig ist.

Da in dieser Verkettung keine Partei zusätzliche Informationen gewinnen soll, spricht man hier auch von einem „Zero-Knowledge Protocol“.

Die Ausarbeitung eines derartigen Zero-Knowledge Protokolls ist die Aufgabe von andern Teilarbeitspaketen; hier soll lediglich veranschaulicht werden auf welchen Arten von Mechanismen ein solches Protokoll funktionieren kann. Im Folgenden werden zwei Beispiele für derartige Mechanismen gegeben:

- Vergleich von Digests von Kennungen statt von Kennungen selbst. Mit gebührenden Vorsichtsmaßnahmen[36] kann dann zwar entschieden werden, ob die Kennungen übereinstimmen, aber ein Empfänger des Digest kann die Kennung selbst nicht ableiten.
- Verschlüsselung einer preisgegebenen Information mit einem Schlüssel, den der Empfänger nur dann errechnen kann, wenn er schon gewisse Information über den Betroffenen kennt. Dies kann offensichtlich vermeiden, Informationen über Betroffene preiszugeben, die kein Konto beim Warndienst haben, und damit auch nicht gewarnt werden können.

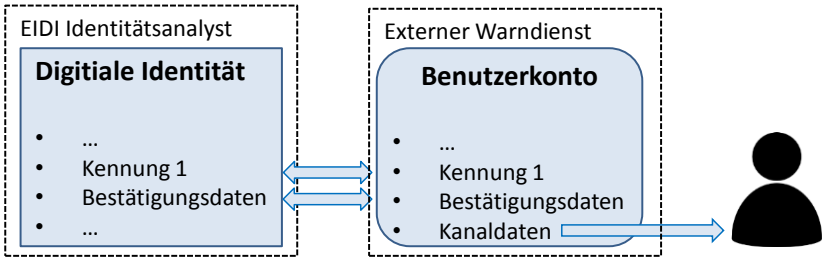

Abbildung 13: Indirekte Warnung aufgrund Verkettung mit externem Benutzerkonto.

8. Anbieter digitaler Identitäten

Die Vorhabensbeschreibung sieht eine Klassierung der Anbieter digitaler Identitäten vor[37]. Durch die sehr breite Definition des Begriffs der digitalen Identität, die für dieses Projekt gewählt wurde (siehe Sektion 2.2 oben), wird auch das Konzept eines Anbieters digitaler Identitäten sehr weit.

Abbildung 2 (siehe Sektion 3 oben) zeigt die Herkunft von digitalen Identitäten nach einem Breach. Damit wäre die Anbieter digitaler Identitäten jegliche Verantwortliche einer ur-

[36] Diese Maßnahmen beinhalten z.B. dass die Inputwerte des Digest-Algorithmus genügend Entropie hat, so dass man die interessanten Digests nicht mit machbarem Aufwand vorberechnen kann. Zu diesen Maßnahmen gehören z.B. die Verwendung von „salt“ oder „nonce“.
[37] Siehe Gesamtvorhabensbeschreibung, 7.3.2, TAP 2.2 Digitale Identitäten, Beschreibung/Ziele des TAP, Seite 14.

sprünglichen Datenverarbeitung. Das Feld dieser Art von Anbietern ist so weit, dass eine nützliche Klassifizierung kaum möglich ist. Im Folgenden werden deshalb nur einige interessante Untermengen von möglichen Anbietern näher betrachtet.

Eine Art von Anbietern sind Betreiber von Online-Diensten, wo Betroffene ein Konto eröffnen können. Ein solches Konto kann durchaus als digitale Teilidentität des Betroffenen betrachtet werden.

Ein wichtiges Unterscheidung-Kriterium für Online-Dienste ist wer zu einem Konto berechtigt ist:

- Öffentlich zugängliche Dienste[38], die offen für alle Interessierten sind,
- Geschlossene Dienste, deren Nutzung auf gewisse Benutzerkreise begrenzt ist, die spezifische Anforderungen erfüllen.

Beispiele für die erste Art von Diensten sind soziale Medien; Beispiel für die letztere Art sind vom Arbeitsgeber oder von einer Vereinigung zur Verfügung gestellte Konten, die nur denjenigen zugänglich sind, die entweder Angestellte oder Mitglieder sind. Die Zugehörigkeit zu einem Dienst beinhaltet dann implizit z.T. weitreichende Information über die betroffenen Personen.

Online-Dienste können nach der Art der von ihnen unterstützten Aktivitäten kategorisiert werden. Abbildung 14 zeigt ein Beispiel, bei dem Sten Franke und das ethority-Team die wichtigsten 250 sozialen Medien 25 verschiedenen Kategorien zugeordnet haben[39].

[38] Kostenpflichtige Dienste können in diesem Sinne auch öffentlich sein.
[39] Quelle: https://ethority.de/social-media-prisma/ heruntergeladen am 18.6.2018.

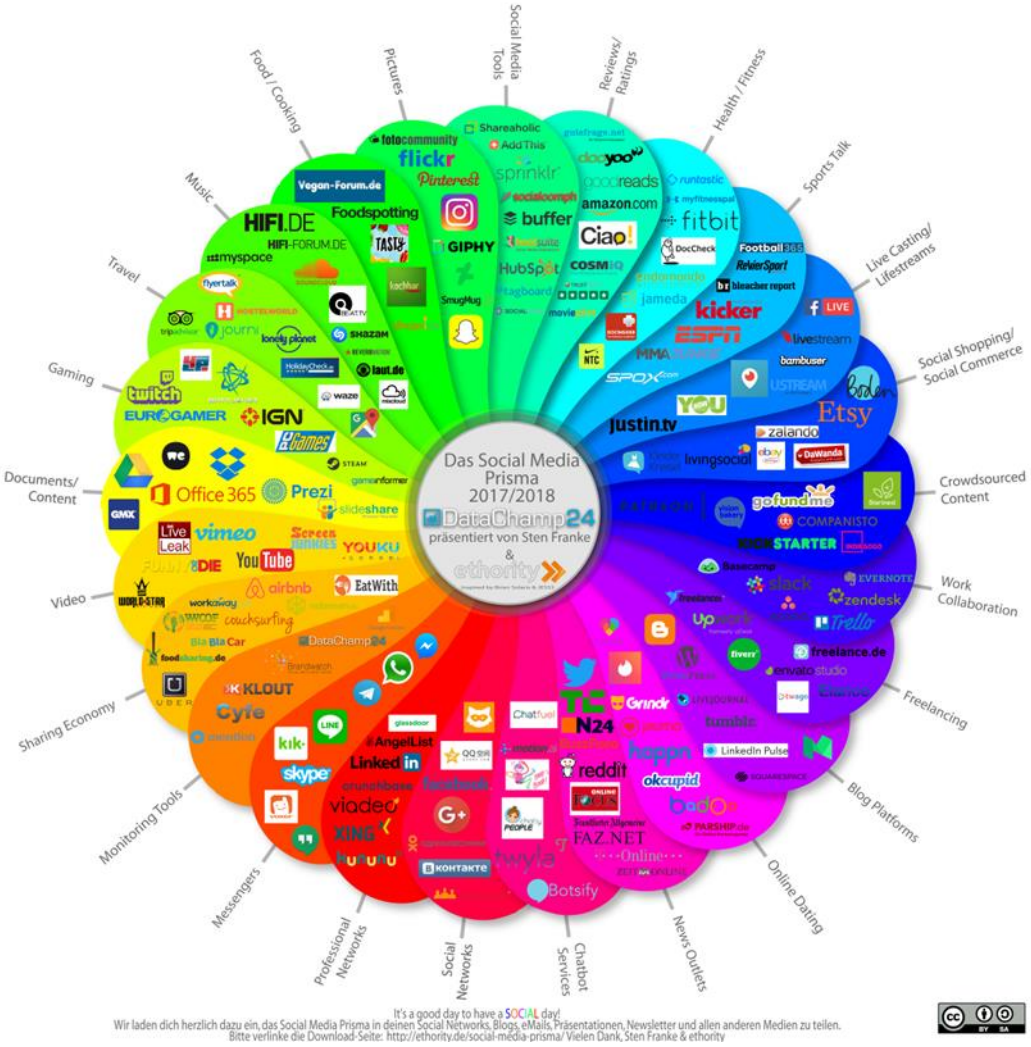

Abbildung 14: Klassierung von Sozialen Medien.

Was diese Kategorisierung und die oben genannten geschlossenen Dienste veranschaulichen ist, dass durch Online-Dienste kreierte digitale Teilidentitäten die betroffene Person immer in einem gegebenen Kontext darstellen. Die bei Online-Diensten übliche digitale Selbstdarstellung und Expression der Betroffenen ist dann spezifisch für diesen Kontext konzipiert und typischerweise auf den Zuhörerkreis der vom Kontext gegeben ist beschränkt. Personen verhalten sich verschieden in verschiedenen Kontexten und normales Verhalten in einem Kontext könnte in einem anderen Kontext als unangebracht bewertet werden. Dies wird veranschaulicht, wenn man die Kontexte von Konversation in einer Partnerschaft, eines Familientreffens, einer Party mit Freunden, und einer beruflichen Umfeld vergleicht. Offensichtlich werden in diesen Kontexten verschiedene Aspekte der Persönlichkeit geteilt.

Diese Erkenntnis ist wichtig im Kontext der Verkettung von Teilidentitäten zwecks Benachrichtigung der Betroffenen. Eine derartige Verkettung kann dann zwar einen Nachrichten-

kanal bereitstellen, der Inhalt einer möglichen Nachricht kann aber stark an die ursprüngliche Teilidentität gebunden sein und im Kontext der Identität des Kanals unerwünscht und unangebracht sein. Ein extremes Beispiel wäre eine Warnung über einen Breach bei einem Online Dating Portal, das an die geschäftliche E-Mail Adresse des Betroffenen geschickt wird.

Idealerweise sollte man dazu alle Anbieter digitaler Identitäten kategorisieren und die Kompatibilität zwischen Kategorisieren einschätzen. Angesichts der Schwierigkeit, die Herkunft geleakter Daten zu bestimmen, und in Betracht der möglichen Anzahl von Kategorien ist ein solches Vorgehen aber praktisch kaum durchführbar.

9. Durch EIDI selbst verursachte Risiken und eine Vorgehensweise zu deren Minimierung

Wie bei jeder Datenverarbeitung entstehen auch bei der Verarbeitung von Breach-Daten durch EIDI Risiken für die Rechte und Freiheiten der Betroffenen. Dies ist in Abbildung 2 rechts durch einen hellblauen Balken dargestellt.

Durch den Forschungscharakter des Projekts sind die von EIDI (also dem Identitätsanalysten) und möglichen Partnern (wie Identitätsdatenwarn- und Informationsdiensten) durchgeführte Verarbeitung nicht vorbestimmt. Was genau wie getan werden kann, um das Gesamtrisiko der von Breaches betroffenen Personen zu verringern, wird erst im Laufe des Projekts erarbeitet. Die von EIDI selbst verursachten Risiken können also erst identifiziert werden, wenn festgelegt wird, welche Daten wie verarbeitet werden. Dies wurde noch durch eine Diskussion von Verkettung von Datensätzen in EIDI und einer Kategorisierung von Anbietern digitaler Identitäten ergänzt.

Beispiele von zu erwartenden Risiken sind die folgenden. EIDI muss für die Erfüllung seines Zwecks geleakte, personenbezogene Daten erfassen. Dadurch ergibt sich das Risiko, dass EIDI selbst zu einer zusätzlichen Datensenke wird, welche die Verfügbarkeit von geleakten Daten erhöht. Abhilfe schaffen offensichtlich Maßnahmen der Vertraulichkeit, die den Zugriff auf diese Daten für andere Zwecke verhindern. Ein weiteres Risiko ist die Kenntnisnahme von personenbezogenen Daten durch EIDI Forscher oder Betreiber des Systems nach abgeschlossenem Projekt. Pseudonymisierung, Zugangskontrolle, und automatische Verarbeitung, die Kenntnisnahme durch Personen unnötig macht, sind Beispiele

für geeignete technische Schutzmaßnahmen; Vertraulichkeitsabkommen und im Vorfeld erarbeitete Vorgehensweisen z.B. wie man reagiert, wenn Forscher Daten einer ihnen bekannten Person zur Kenntnis nehmen, sind Beispiele für organisatorische Maßnahmen.

Offensichtlich können die konkreten Risiken und die dazu passenden Maßnahmen nur im Detail bestimmt werden, ab dem Zeitpunkt, wo mögliche Verarbeitungsschritte diskutiert werden. EIDI hat genau dafür das Teilarbeitspaket 5.2 *Datenschutzrechtliche Begleitung* eingeplant. Diese Vorgehensweise ermöglicht es, sich schon vor der Verarbeitung von Daten der Risiken bewusst zu werden und geeignete Schutzmaßnahmen zur Beschränkung der Risiken zu treffen. Dies steht im Einklang mit dem Prinzip des Datenschutzen durch Technikgestaltung („data protection by design") und der Vorgehensweise einer Datenschutzfolgenabschätzung[40].

10. Fazit

Das vorliegende Deliverable hat mit der Definition des Begriffs der digitalen Identität angefangen und dann aufgrund einer Modellierung von EIDI die datenschutzrelevanten Merkmale identifiziert. Diese wurden dann im Detail analysiert. Sowohl das Modell als auch die Analyse der Merkmale unterstützen zukünftige Entscheidungen im Projekt, welche Verarbeitung am besten dazu geeignet ist, die Risiken der Betroffenen nach einem Breach zu verringern.

[40] Siehe Art. 35 der DSGVO.