Übersicht Anonymität

Ein Ausfluss des weltweiten Digitalisierungsprozesses ist die Entstehung der Opendata- und Opengovernment-Movements. Diese haben zum Ziel, Daten der Öffentlichkeit zur Verfügung zu stellen und somit externen Wissenschafter*innen, Journalist*innen, und Bürger*innen die Mittel bieten, Hypothesen zu verifizieren oder explorative Datenanalyse zu betreiben. Dies sorgt für die Zuordnung der Verantwortung von Regierungen und kann somit die Demokratie stärken. Die Behörden des Kantons haben diesen Auftrag (Art. 26 BVE).
Jedoch haben die Entitäten, welche Daten mit personenbezogenen Angaben besitzen, eine Pflicht, die Daten Missbrauch zu schützen (Art. 13 Abs. 2). Bei der Veröffentlichung von Daten wird dies durch Anonymisierung erreicht.
Die Stärke der Anonymisierung basiert jedoch immer auf einer Interessenabwägung zwischen dem öffentlichen Interesse an neuen Erkenntnissen und dem Schutz der von den Daten betroffenen Personen.

Identifikation: Eindeutige Zuordnung einer Information zu einer natürlichen Person.

Anonymisierung: Entkopplung einer Information von einer natürlichen Person, in dem Masse, dass diese nicht mehr identifiziert werden kann.

Identifikator: Angaben, anhand welcher eine Person eindeutig identifizieren

Pseudonymisierung: Ersetzen von personenbezogenen Identifikatoren mit Angaben, die ohne Schlüssel nicht einer Person zugewiesen werden können.

Quasi-Identifizierer: Spalten, welche zur Identifikation benutzt werden können, aber selbst noch nicht zur Identifikation reichen (z.B. Adressen)

Sensitive Attribute: Spalten, welche nicht mit einem Individuum in Verbindung gebracht werden sollten (z.B. Krankheiten).

Äquivalenzklasse: Eine Menge von Zeilen, bei der die Quasi-Identifizierer jeweils identische Werte haben.

k-Anonymität Idee: Kombinieren der Zeilen mit gleichen Sensitiven Attributen bis alle vorkommenden Quasi-Identifizierenden Werte mindestens k mal vorhanden sind.
k-Anonymität alleine ist oft nicht hinreichend (siehe weiterführende Links).

ℓ-Diveristät: k-Anonym, mit der zusätzlichen Anforderung, dass mindestens ℓ verschiedene Sensitive Attribute für jede Äquivalenzklasse vorhanden sind.
Garantiert, dass wenn beliebig viele der Quasi-Identifizierer einem "Angreifer" bekannt sind immer noch ℓ viele Sensitive Attribute in Frage kommen (z.B. wenn ℓ=2 in und die Sensitiven Attribute Krankheiten Sind kann der Angreifer evtl. nicht wissen ob jemand Herzprobleme oder Tuberkulose hatte).

Trade Offs: Bei höheren Werten für ℓ oder k geht mehr Information verloren, da eben Zeilen zusammengelegt werden müssen.
Wie in der Einleitung erwähnt, gibt es eine Balance zu halten zwischen der Nutzbarkeit der Daten (Öffentliches Interesse) und dem Anonymitäts Anspruch des Individuums.

Differential Privacy: In diesem mathematischen Framework werden Anonymität und Verlust der Information parametrisiert, gegeneinander aufgewogen und parametrisiert. Dazu werden die Daten nicht herausgegeben, sondern Abfragen nach Daten mit geschickt gewählten, zufalls behafteten Daten beantwortet, sodass der Fehler in Statistiken sehr klein ist, jedoch keine Rückschlüsse auf Individuen gezogen werden können.
Es gibt auch Software Frameworks, welche bei der Implementierung von APIs, welche die mathematischen Ansprüche erfüllen, helfen (Siehe Weiterführende Links).

Werkzeuge:

Open Source GUI Tool für messen der Anonymität in Datensätzen:
https://arx.deidentifier.org/

Python Bibliothek und Papier zu Massen von Datensatz
https://github.com/IFCA/pycanon

Weiterführende Links

PDF zu Anonymität der EU:
https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf

Zum Trade off von ℓ und den verbleibenden Daten:
https://cloud.google.com/dlp/docs/compute-l-diversity

Beispiel für die schwäche der k-anonymität:
https://en.wikipedia.org/wiki/K-anonymity#Possible_attacks

Einführungen in differential privacy:
https://de.wikipedia.org/wiki/Differential_Privacy
https://desfontain.es/privacy/friendly-intro-to-differential-privacy.html

Buch zu differential privacy:
The Algorithmic Foundations of Differential Privacy

Bibliothek zur implementierung von differential privacy:
https://github.com/google/differential-privacy