Meta-data about anonymised data as a means to enable the future data economy

Benjamin Heitmann (Ph.D.)

RWTH Aachen, Informatik 5
Lehrstuhl Prof. Decker

Fraunhofer Institute for Applied Information Technology FIT

Privacy as an enabler of the data economy

- To enable the data economy, selling data has to be enabled
- One way to enable this is anonymisation

- Anonymisation is very well established by now
- **Anonymisation changes the utility of the data!**
- **The utility corresponds to the value of the data in a market place**

- The time is right to standardise the details of how data is anonymised.

- **Future perspective:**
  - Similar standardisation will be necessary for encrypted data in the future
  - Data protected by the **blockchain** is opaque
  - So is data encrypted with Homomorphic Encryption and similar cryptographic **PETs**

Important attribute subsets for tabular data

- Key attributes, also called **Identifiable Attributes (IA)**
  - Attributes directly reveling the identity of tuple
  - Full name, phone number, SSN etc.
  - Always suppressed (removed) before release!

- **Quasi-identifiers (QID)**
  - Set of Attributes which can revel the identity of a tuple!
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.

- **Sensitive Attributes (SA)**
  - Attributes which express some information of the tuple
  - Type of illness, reason for arrest, grades etc.
  - These attributes is what the researchers need

- These three sets are disjoint!

K-anonymity with k = 2

Name (IA)	Age (QID)	Gender (QID)	Zip (QID)	Disease (SA)
*	30-40	male	520**	Flu
*	30-50	male	520**	Burnings
*	20-30	female	520**	Stab wound
*	30-40	male	520**	Gun shot
*	20-30	female	520**	HIV
*	30-50	male	520**	Blind

Name (IA)	Age (QID)	Gender (QID)	Zip (QID)
Ahri	27	female	52064
Annie	16	female	52061
Ashe	30	female	52075
Brand	34	male	52080
Camille	27	female	52073
Lucian	34	male	52077
Ezreal	20	male	52064

- Re-identification rate drops to 1/k

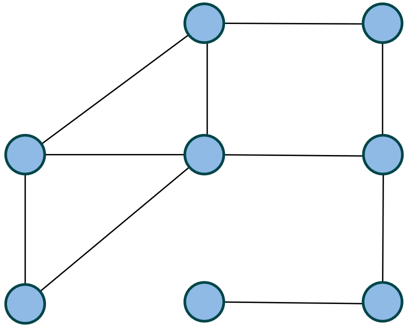
More information on Re-identification by Liking
Latanya Sweeney's original k-anonymity paper (1997)

l-Diversity

Age (QID)	Salary (QID)	Zip (QID)	Disease (SA)
2*	20K	476**	Gastric Ulcer
2*	30K	476**	Gastritis
2*	40K	476**	Stomach Cancer
≥40	50K	4790*	Gastritis
≥40	100K	4790*	Flu
≥40	70K	4790*	Bronchitis
3*	60K	476**	Bronchitis
3*	80K	476**	Pneumonia
3*	90K	476**	Stomach Cancer

Anonymisation of graph data: K-degree Anonymity for k=2

Input graph:



Example for modifying a graph:

- greedy edge addition

$$d = \begin{pmatrix}4\\3\\3\\3\\2\\2\\2\\1\end{pmatrix} \qquad d' = \begin{pmatrix}4\\4\\4\\4\\2\\2\\2\\2\end{pmatrix} \qquad d' - d = \begin{pmatrix}0\\1\\1\\1\\0\\0\\0\\1\end{pmatrix}$$