Anonymization of Micro Data

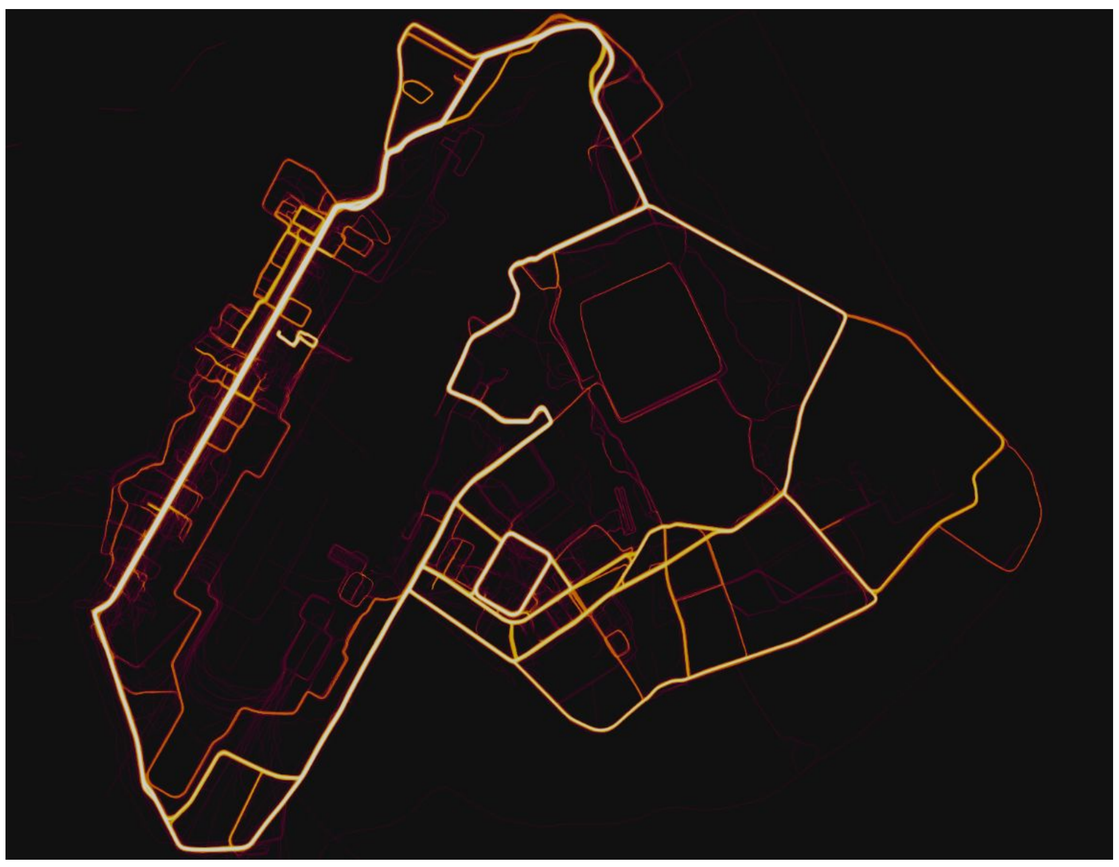
Techniques and Attacks

Releasing confidential user data to the public is often done because of requirements by laws or to provide researchers with data sets. Journalists can also find evidence in data releases to support stories, e.g., payments to doctors by pharma companies [1]. To protect the users' privacy, it should not be possible to link any information in a data release to real people or records in other databases. If, however, too much information is removed, the published data set loses its usefulness.

Micro Data

Anonymization is especially difficult, when publishing micro data, i.e., data where each record represents a real person. Even from partially anonymized and redacted data it may be possible to find out where a target person lives or works [2]. The techniques explained on the right mitigate these issues, but are still not attack-proof.

Attacks despite anonymization



Activity heatmap in Bagram Air Base (Afghanistan) leaked from fitness tracking data

Fitness tracking company Strava released a global heatmap of location traces obtained from its users. Although user-defined privacy zones and all identifiers have been removed [6], routes taken by soldiers inside military bases can be identified [7]. If published as micro data, even more information could have been revealed.

k-Anonymity

The data is separated into **quasi-identifiers**, which may be used to link with external data and **sensitive attributes**. Quasi-identifiers are generalized, such that each cluster with identical attributes contains at least *k* records [3].

Age	City	Disease
25	Oslo	Flu
27	Oslo	None
33	Bern	Cancer
37	Bern	Cancer

Original data

Age	City	Disease
20-29	Oslo	Flu
20-29	Oslo	None
30-39	Bern	Cancer
30-39	Bern	Cancer

cancer inferred

2-anonymized

Homogeneity Attack: If a sensitive attribute is identical for all records in a cluster, the sensible attribute can still be inferred.

ℓ-Diversity

To prevent a homogeneity attack, the clusters must also contain at least ℓ different values for each sensitive attribute [4].

Age	City	Disease
25	Oslo	Flu
27	Oslo	None
29	Oslo	Allergy
33	Bern	Cancer
37	Bern	Cancer
39	Bern	None

Original data

Age	City	Disease
20-29	Oslo	Flu
20-29	Oslo	None
20-29	Oslo	Allergy
30-39	Bern	Cancer
30-39	Bern	Cancer
30-39	Bern	None

high cancer chance

3-anonymized, 2-diverse

Skewness Attack: If the relative frequency of a value within a cluster differs wildly from the overall one, a possibly more sensitive value can be strongly predicted for a target.

Conclusion

To prevent attacks on ℓ-diverse data releases there are additional measures modifying sensitive attributes like *t*-closeness [5]. As seen in the Strava example however, careful consideration is always required, even when adhering to all principles or with all identifiers removed, as data complexity is expected to grow in the future.

Simon Krogmann
Master Student
Hasso Plattner Institute,
Potsdam, Germany
simon.krogmann@student.hpi.de

Poster for *Data Engineering Lecture Series*

References:

[1] S. Wehrmeyer. *Data Engineering im Newsroom.* Presentation, Data Engineering Lecture Series, 2017.
[2] L. Heuckendorf. *The Science behind Visual Analytics.* Presentation, Data Engineering Lecture Series, 2017.
[3] P. Samarati, and L. Sweeney. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* Technical report, SRI International, 1998.
[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. *l-diversity: Privacy beyond k-anonymity.* Data Engineering, 2006.
[5] N. Li, T. Li, and S. Venkatasubramanian. *t-closeness: Privacy beyond k-anonymity and l-diversity.* Data Engineering, 2007.
[6] D. Robb. *The Global Heatmap, Now 6x Hotter.* Blog article, Medium, 2017.
[7] BBC News. *Fitness app Strava lights up staff at military bases.* News article, BBC, 2018.

Hasso Plattner Institut