Why Open Data May Threaten Your Privacy

Sebastian Pape, Jetzabel Serna-Olvera, Welderufael B. Tesfay

Goethe University Frankfurt
Chair of Mobile Business and Multilateral Security

Workshop on Privacy and Inference
September 21st, 2015

Overview

1 Open Data / De-Anonymization

2 Proposed Approach

3 Conclusion

Open Data / De-Anonymization

Open Data

- Broad range of Applications
- Services helpful to society (e.g. health, educational services)
- Balancing act between usefulness and anonymization

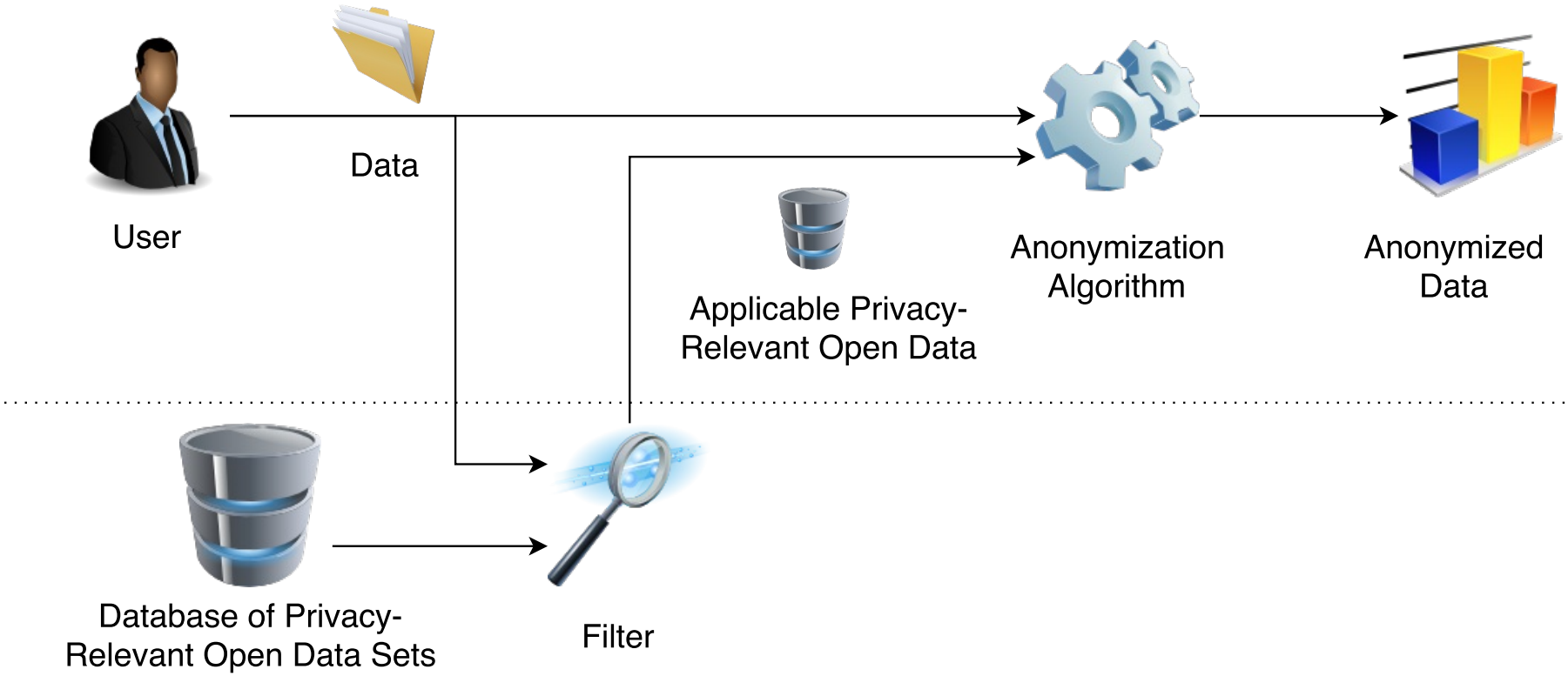
De-Anonymization

- Often works by linking data sets “unexpectedly”
- Gets easier with more Open Data
- Machine Learning allows to work with fuzzy data

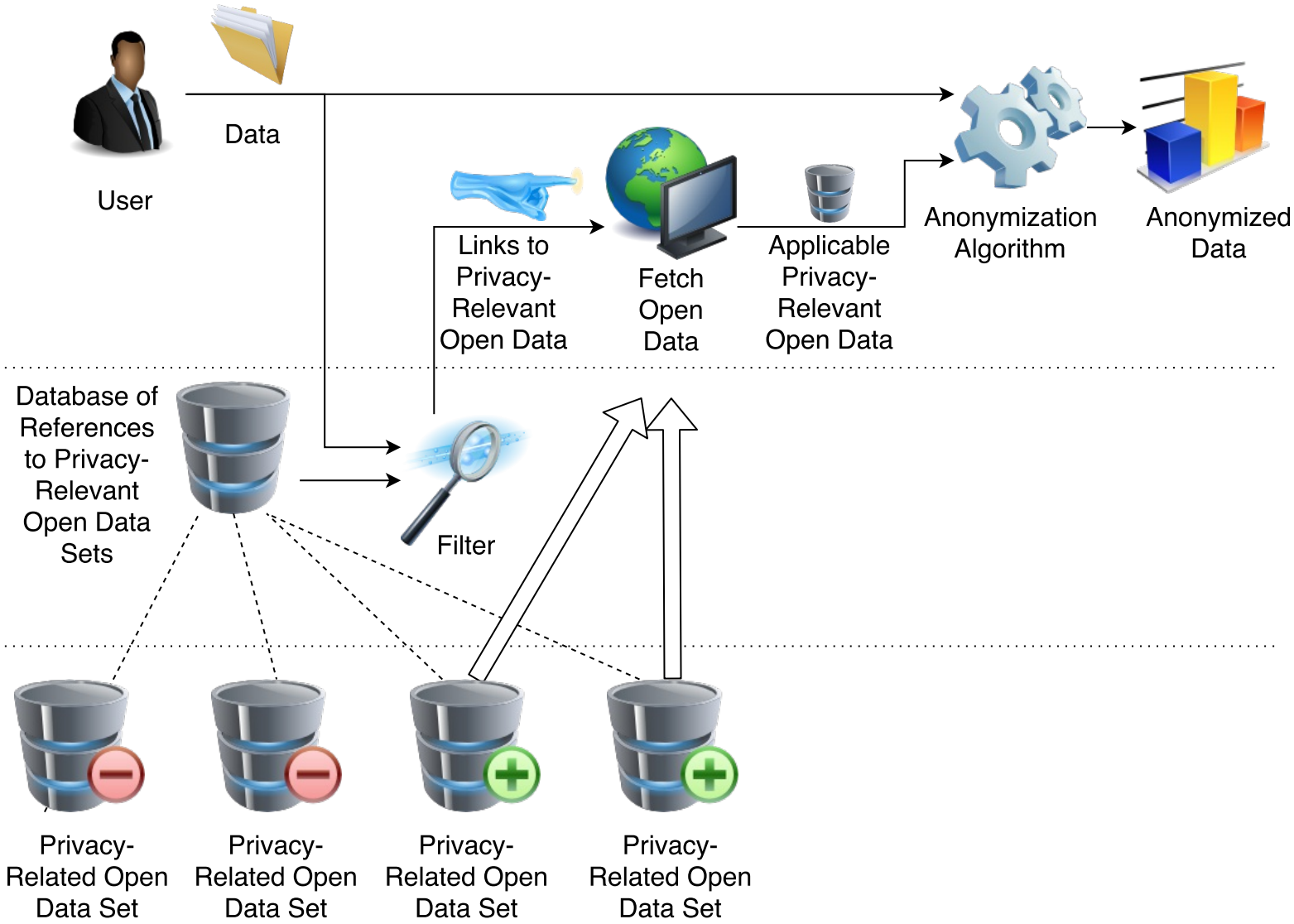
Tool Support

- Not the fault of anonymization algorithms
- Tool support to identify relevant Open Data needed
- Several capabilities for machine learning approaches
- Scope limited to Open Data

Mirroring Privacy-Related Open Data

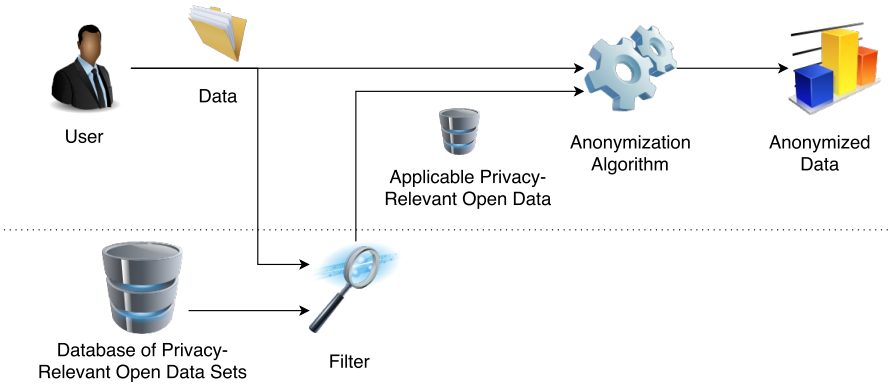


Referencing Privacy-Related Open Data



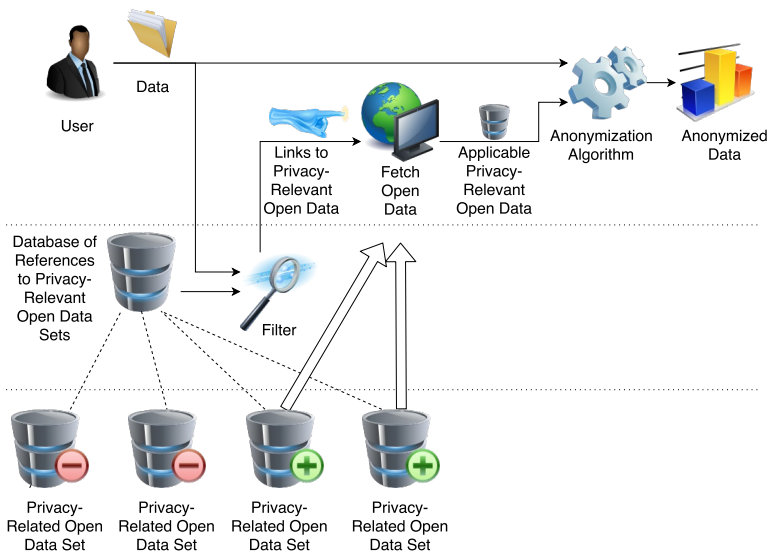
Comparison

Mirroring



+ Usability
+ Quality of Prediction
+ Versioning
- Updates
- Storage
- Bandwidth

Referencing



- Usability
= Quality of Prediction
- Versioning
= Updates
+ Storage
= Bandwidth

Proposed Steps

Steps

- C Collection (O)
- L Linkage (T)
- U User Interaction (O/T)
- A Anonymization as a Service (T)
- D De-Anonymization-Tests as a Service (T)

Challenges

Steps

- C Collection (O)
- L Linkage (T)
- U User Interaction (O/T)
- A Anonymization as a Service (T)
- D De-Anonymization-Tests as a Service (T)

Challenges

- C1 Rate Privacy-Relevance
- C2 Version Control System
- L1 Context of Database
- L2 Field Names
  - DCAT, VoID
- L3 Sparse matches
- U1 Structure vs. Full Data
- A1 Server Side vs. Client Side Analysis
- D1 Deterministic vs. Probabilistic Model

Conclusion / Discussion

- Open Data: More attention should be paid to privacy risks
- Tool support needs to be improved
- Publication of Open Data should not be prevented
  - Ohm (2009) vs. O'Hara (2011)
- Useful tool or threat?
  - → Short-term: threat
  - → Long-term: useful tool
- Should the tool regard leaked/stolen data?

Feedback

sebastian.pape

jetzabel.serna

welderufael.tesfay

@m-chair.de