RECORD LINKAGE AND PPRL 
with Clustering of Matches

Erhard Rahm, Martin Franke
AGENDA

- Record Linkage with Clustering
  - FAMER Tool

- Privacy-Preserving Record Linkage (PPRL)
  - PRIMAT Tool

- Summary
- input: 1, 2 or n data sources
- clustering can improve match quality
  - additional matches
  - removal of wrong matches
- compact representation of matches
  - cluster with k records corresponds to $k(k-1)/2$ match pairs
  - e.g. 10 elements instead of 45 match pairs
FAMER TOOL

**FAst Multi-source Entity Resolution System**
- scalable linking & clustering for many sources

---

**Input**
- Source A
- Source B
- Source D
- Source C
- Source E

**Linking:** Similarity Graph

**Clustering**
INCREMENTAL MATCHING & CLUSTERING

Data sources (continuously changing)

Stream of new entities

Entities of new source

Integrated knowledge graph

Cluster Set

DS 1

DS 2

DS n

new DS
AGENDA

- Record Linkage mit Clustering
  - FAMER Tool

- Privacy-Preserving Record Linkage (PPRL)
  - PRIMAT Tool

- Zusammenfassung
PPRL REQUIREMENTS

- high degree of privacy
  - no forwarding of unencoded quasi-identifiers to other institutions
  - minimal attack risk, e.g. no forwarding of person data to multiple sites or no central storage of (encoded) patient lists

- High performance and scalability
  - Many institutions, many persons/patients

- high match quality on encoded data
  - robustness against data errors/deviations, e.g. changed last name
  - support for match groups (clusters), not only match pairs

- dynamic PPRL: support for additional patients/ institutions

- flexible usability in different projects / scenarios
  - possible need to allow re-identification of patients, e.g. for recruitment in new clinical study

- suitable tool support
  - flexible configurability, support for different usage forms
  - ease of use
BATCH VS. INCREMENTAL MATCHING

- **Batch Matching**
  - linkage on fixed set of data
  - no storage of data and match results at linkage unit
  - re-computation needed for changed data

- **Incremental Matching**
  - encoded person records and matches are stored in database
  - incremental matching for new records
  - faster compared to complete re-computation
PRIMAT TOOLBOX

- Private Matching Toolbox (Uni Leipzig)
- open-source PPRL-Tool for entire PPRL process
  - components for data owner and Linkage Unit
  - batch or incremental matching
- flexible configuration and execution of PPRL workflows
- high performance by blocking and parallel matching

Franke, M.; Sehili, Z.; Rahm, E.: PRIMAT: A Toolbox for Fast Privacy-preserving Matching. PVLDB 2019
## COMPARISON OF PPRL TOOLS

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mainzelliste (Lablans et. al.)</th>
<th>PRIMAT (U Leipzig)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open source</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>data cleaning</td>
<td>(✔)</td>
<td>✔</td>
</tr>
<tr>
<td>Flexible encodings (Bloom Filter)</td>
<td>field-level Bloom Filter</td>
<td>✔</td>
</tr>
<tr>
<td>Hardening support</td>
<td>✘</td>
<td>✔</td>
</tr>
<tr>
<td>private blocking (with LSH)</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Parallel matching</td>
<td>✘</td>
<td>(✔)</td>
</tr>
<tr>
<td>Match grouping / clustering</td>
<td>without transitivity</td>
<td>✔</td>
</tr>
<tr>
<td>Central or decentralized linkage</td>
<td>central</td>
<td>central</td>
</tr>
<tr>
<td>Batch + incremental Matching</td>
<td>Incremental</td>
<td>(✔)</td>
</tr>
</tbody>
</table>
CLUSTERING IN MAINZELLISTE

- Mainzelliste tool used in many projects in Germany
- Cluster is represented by oldest (first) record
- New record is assigned to cluster with highest similarity (above threshold)
- Inserts are processed one record at a time (slow, results can depend on order of inserts)

Problems:

- Poor quality of cluster representant (first record) can reduce match quality
- Fixed cluster representant cannot deal with changes in address, name etc.
- Joint insertion of multiple records or entire sources can improve performance and quality
CLUSTERING MAINZELLISTE
PROBLEM SCENARIO

Match between records 2 and 3 is not found since record 3 is only compared with record 1.
CLUSTERING MAINZELLISTE
PROBLEM II

record-wise matching

joint consideration of multiple records
From duplicate-free source
CLUSTERING EXTENSIONS

- Record for sources whether they are duplicate-free
  - optimized cluster assignment (per cluster at most one record of duplicate-free source)

- Comparison with all cluster members or only 1 representant
  - trade-off between quality and scalability

- Dynamic selection of cluster representant
  - youngest (newest) record
  - record with highest average similarity to all other cluster members
  - 'virtual' cluster representant by aggregating cluster members, e.g. with Counting Bloom Filter
COUNTING-BLOOM-FILTER AS DYNAMIC CLUSTER REPRESENTANT

Max Muster 07.10.1969 Ritterstraße 9

| 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |

Max Muster 07.10.1969 Humboldstraße 25

| 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |

Sim(r2, CBF) = 3/5 = 0.6
Sim(r2, r1) = 3/5 = 0.6

Max Meier 07.10.1969 Humboldstraße 25

| 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |

Sim(r3, CBF) = 3/6 = 0.5
Sim(r3, r1) = 2/6 = 0.33
Sim(r3, r2) = 3/5 = 0.6
SUMMARY

- Record Linkage: pipeline with blocking, matching and clustering
- PPRL: many, partially project-specific requirements
- Flexible configurability needed
  - batch and/or incremental matching
  - preprocessing, encoding, blocking, matching, clustering
- Current PPRL tools provide no or insufficient support for clustering
  - Need grows with more data sources and more possible duplicates
- PRIMAT: configurable toolbox with high flexibility and performance
REFERENCES RECORD LINKAGE

- A. Saeedi, E. Peukert, E. Rahm: *Using Link Features for Entity Clustering in Knowledge Graphs*. Proc. ESWC 2018 (Best research paper award)
REFERENCES PPRL

- M. Franke et al.: *Post-processing Methods for High Quality Privacy-Preserving Record Linkage*. Proc, DPM, LNCS 2018
- M. Franke et al.: *ScaDS Research on Scalable Privacy-preserving Record Linkage*. Datenbank-Spektrum 2019