ScaDS DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

Privacy-preserving data integration for Big Data

Martin Franke, Eric Peukert, Ziad Sehili, Erhard Rahm

PRIVACY

Privacy

- right of individuals to determine by themselves when, how and to what extent information about them is communicated to others (Agrawal 2002)

Privacy threats

- extensive collection of personal/private information / surveillance
- Information dissemination: disclosure of sensitive/confidential information
- Invasions of privacy: intrusion attacks to obtain access to private information
- Information aggregation: combining data, e.g. to enhance personal profiles or identify persons (de-anonymization)

INFORMATION / DATA PRIVACY

- protection especially critical for *personally identifiable information* (PID), also called quasi-identifiers
  - name, birthdate, address, email address etc
  - healthcare and genetic records, financial records
- challenge: preserve privacy despite need to use person-related data for improved analysis / business success (advertisement, recommendations), website optimizations, clinical/health studies, identification of criminals …
  - tracking and profiling of web / smartphone / social network users (different kinds of cookies, canvas fingerprinting …)
  - often user agreement needed

ScaDS DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

A privacy reminder from Google

To be consistent with data protection laws, we're asking you to take a moment to review key points of Google's Privacy Policy. This is not about a change we've made - but please review the key points below. **Click "I agree" to agree to the terms set out below; you can also explore other options on this page.** You can revoke your consent at any time with effect for the future.

Usage and content data

- When you use Google services to do things like write a message in Gmail or comment on a YouTube video, we store the information you create.
- When you search for a restaurant on Google Maps or watch a video on YouTube, for example, we process information about that activity — including information like the video you watched, device IDs, IP addresses, cookie data, and location.
- Our Privacy Policy contains further descriptions of the data we process.
- We treat all of this as "personal information" when it's associated with your Google Account.
- We also process the kinds of information described above when you use apps or sites that use Google services like ads, Analytics, and the YouTube video player.

https://www.google.com/intl/en/policies/privacy/

ScaDS DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

Google Information that we collect

- **Information you give us.** For example, many of our services require you to sign up for a Google Account. When you do, we'll ask for personal information, like your name, email address, telephone number or credit card to store with your account. If you want to take full advantage of the sharing features we offer, we might also ask you to create a publicly visible Google Profile, which may include your name and photo.
- **Information we get from your use of our services.** We collect information about the services that
  - Device information

    We collect device-specific information (such as your hardware model, operating system version, unique device identifiers, and mobile network information including phone number). Google may associate your device identifiers or phone number with your Google Account.
  - **Log information**
  - **Location information**

    When you use Google services, we may collect and process information about your actual location. We use various technologies to determine location, including IP address, GPS and other sensors that may, for example, provide Google with information on nearby devices, Wi-Fi access points and mobile towers.
  - **Unique application numbers**

ScaDS DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

Google

Purposes of the data processing

We process this data for the purposes described in our policy, including to:

- Help our services deliver more useful, customized content such as more relevant search results, based on your interests derived from such data;
- Improve the quality of our services and develop new ones;
- Deliver ads based on your interests, which we can determine based on this data, like ads that are related to things such as search queries or videos you've watched on YouTube;
- Improve security by protecting against fraud and abuse; and
- Conduct analytics and measurement to understand how our services are used.

Combining data

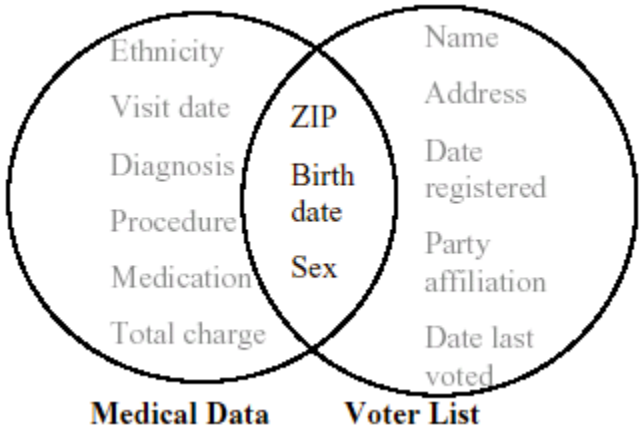
We also combine this data among our services and across your devices for these purposes. For example, we show you ads based on information about your interests, which we can derive from your use of Search and Gmail, and we use data from trillions of search queries to build spell-correction models that we use across all of our services.

PRIVACY FOR BIG DATA

- need for comprehensive privacy support (“privacy by design”)
- privacy-preserving publishing of datasets
  - anonymization of datasets
- privacy-preserving data mining
  - analysis of anonymized data without re-identification
- privacy-preserving record linkage
  - object matching with encoded data to preserve privacy
  - prerequisite for privacy-preserving data mining

RE-IDENTIFICATION OF „ANONYMOUS DATA“ (SWEENEY 2001)

- US voter registration data
  - 69% unique on postal code (ZIP) and birth date
  - 87% US-wide with sex, postal code and birth data



- Solution approach: K-Anonymity
  - any combination of values appears at least k times
  - generalize values, e.g., on ZIP or birth date

K-ANONYMITY EXAMPLE

ID	ZIP	AGE	DISEASE	TREATMENT
1	12345	23	Gastric ulcer	Antacid
2	12345	29	Gastritis	Acid-reducing drug
3	12363	41	Flu	Antipyretic drug
4	12361	43	Stomach cancer	Cytostatic drug
5	12362	59	Pneumonia	Antibiotics
6	12471	52	Bronchitis	Antibiotics
7	12473	55	Flu	Antipyretic drug

(a) Microdata-table

ID	ZIP	AGE	DISEASE	TREATMENT
1	123**	[20-29]	Gastric ulcer	Antacid
2	123**	[20-29]	Gastritis	Acid-reducing drug
3	123**	[40-49]	Flu	Antipyretic drug
4	123**	[40-49]	Stomach cancer	Cytostatic drug
5	123**	[50-59]	Pneumonia	Antibiotics
6	124**	[50-59]	Bronchitis	Antibiotics
7	124**	[50-59]	Flu	Antipyretic drug

(b) 2-anonymous table

from: Nielsen et al: Proc BTW 2015

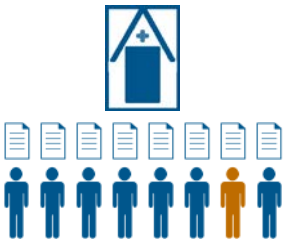
ANONYMIZATION VS PSEUDONYMIZATION

- Anonymization
  - removing, generalizing or changing personally identifying attributes so that people whom the data describe remain anonymous
  - different records for same person cannot be matched/combined
- Pseudonymization
  - quasi-identifiers are replaced by one or more artificial identifiers (pseudonyms)
  - one-way pseudonymization (e.g. one-way hash functions) vs. re-identifiable pseudonymization
  - records with same pseudonym can be matched
  - improved potential for data analysis

ScaDS DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

PRIVACY-PRESERVING RECORD LINKAGE (PPRL)



Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges

Dinusha Vatsalan, Ziad Sehili, Peter Christen and Erhard Rahm

Abstract The growth of Big Data, especially personal data dispersed in multiple data sources, presents enormous opportunities and insights for businesses to explore and leverage the value of linked and integrated data. However, privacy concerns impede sharing or exchanging data for linkage across different organizations. Privacy-preserving record linkage (PPRL) aims to address this problem by identifying and linking records that correspond to the same real-world entity across several data sources held by different parties without revealing any sensitive information about these entities. PPRL is increasingly being required in many real-world application areas. Examples range from public health surveillance to crime and fraud detection, and national security. PPRL for Big Data poses several challenges, with the three major ones being (1) scalability to multiple large databases, due to their massive

PRIVACY-PRESERVING RECORD LINKAGE (PPRL)

- record linkage / object matching with encoded data to preserve privacy
  - data exchange / integration of person-related data
- privacy aspects
  - need to support secure 1-way encoding (pseudonymization)
  - protection against attacks to identify persons
- conflicting requirements:
  - high privacy
  - match effectiveness (need to support fuzzy matches)
  - scalability to large datasets and many parties

PPRL USE CASES

- medical domain (patient data)
  - central registry for certain diseases, e.g. cancer
  - clinical studies to optimize treatments based on combined data from several hospitals, physicians, etc.
  - protected combination of medical data with other data sources (e.g., on unemployment, migration, ...) for social studies
- criminalistics
  - protected combination of information from banks, credit card companies, email service providers, etc. for suspicious persons
  - detection of criminal merchants in online shops / dark net

...

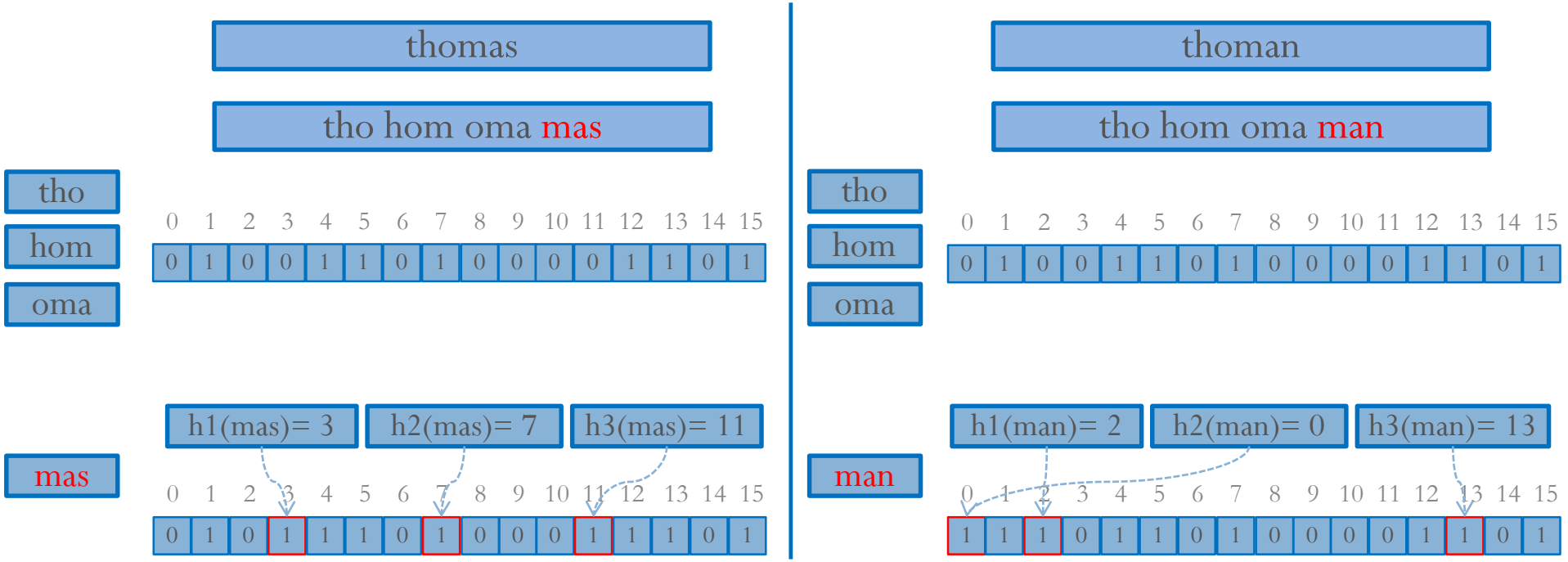
PPRL APPROACHES

- data encoding
  - bloom filters, embeddings (mapping to points in metric space), cryptographic encryption, ...
- involved parties
  - two or more data owners
  - central linkage unit (LU) or symmetric protocol
- privacy model
  - *honest-but-curious* vs *malicious* parties
  - considered types of attacks (frequency, dictionary, collusion, ...)
- blocking and matching approaches for encoded data

PPRL WITH BLOOM FILTERS

- effective and simple encoding uses cryptographic bloom filters (Schnell et al., 2009)
- tokenize match-relevant quasi-identifiers, e.g. using bigrams or trigrams
  - typical attributes: first name, last name (at birth), sex, date of birth, country of birth, place of birth
- map each token with a family of one-way hash functions to fixed-size bit vector (fingerprint)
  - original data cannot be reconstructed
- match of bit vectors (e.g., using Jaccard similarity) is good approximation of true match result

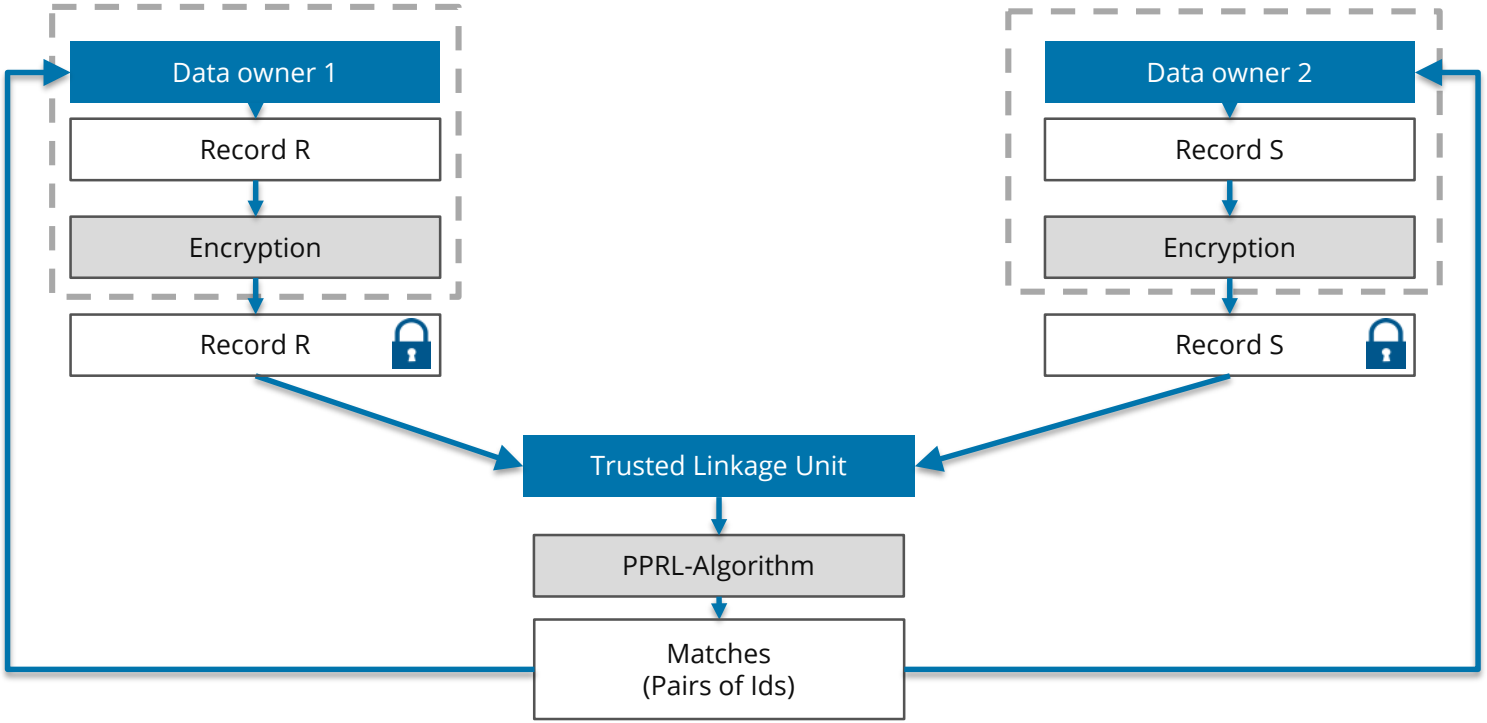
SIMILARITY COMPUTATION – EXAMPLE

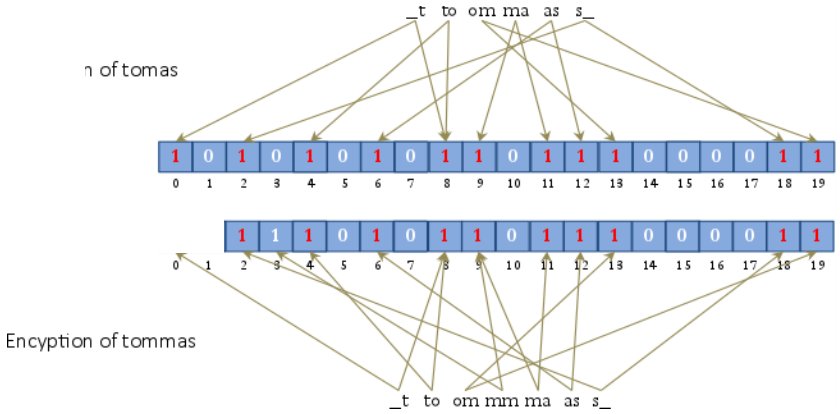


$$Sim_{Jaccard}(r1, r2) = (r1 \wedge r2) / (r1 \vee r2)$$

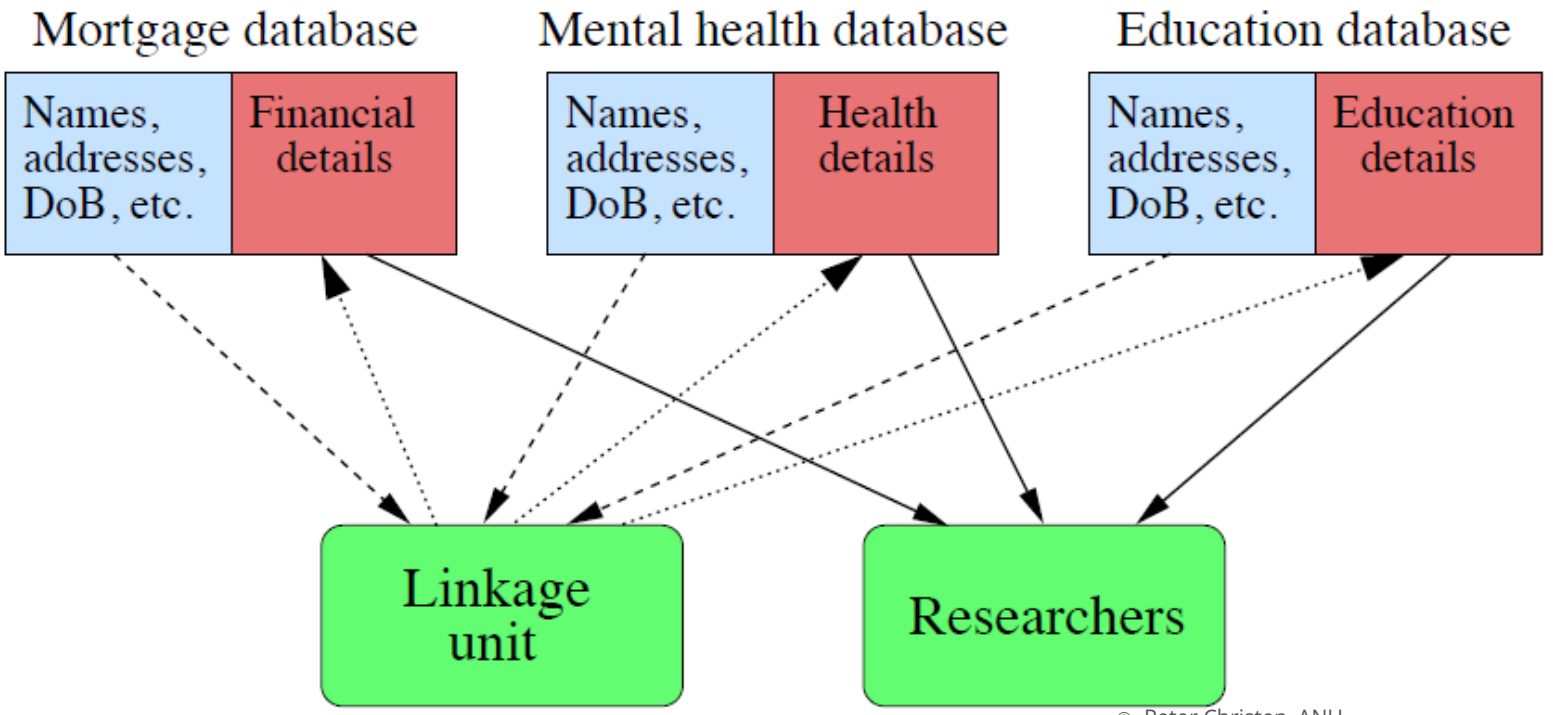
$$Sim_{Jaccard}(r1, r2) = 7/11$$

PPRL PROTOCOL



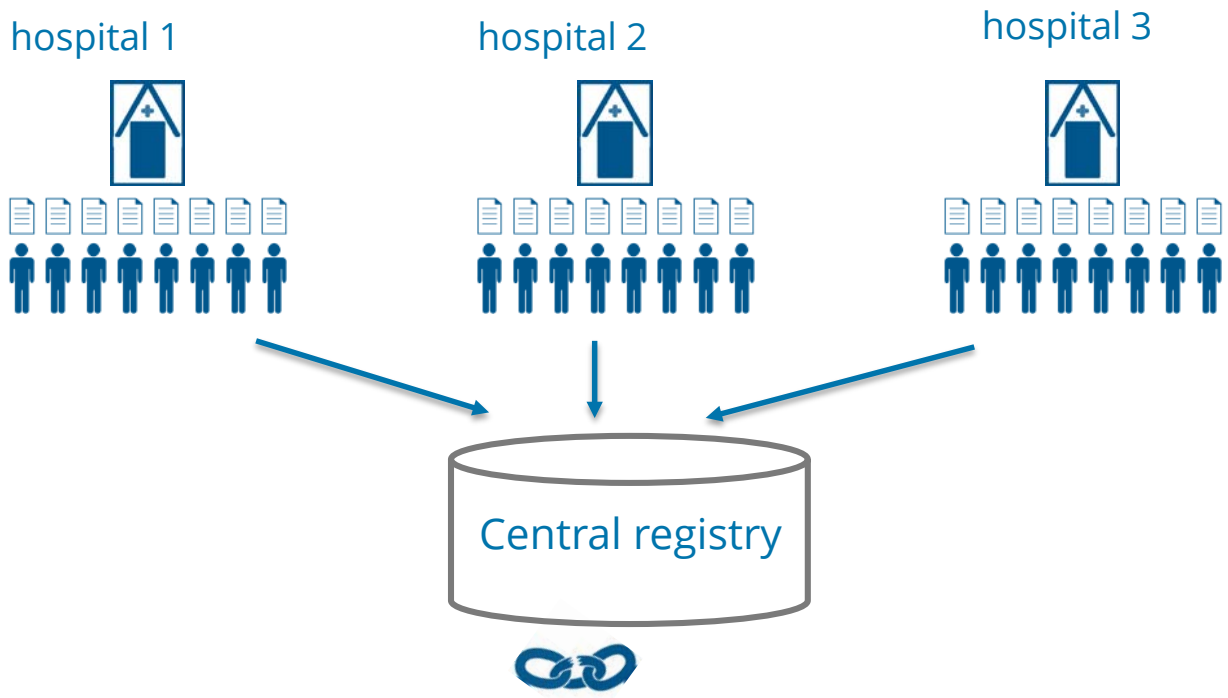


SAMPLE PPRL USE CASE (1)



C.W. Kelman, A.J. Bass. C.D. Holman: Research use of linked health data--a best practice protocol. Aust NZ J Public Health. 2002

SAMPLE PPRL USE CASE (2)



PPRL linking to identify
and eliminate duplicates

SCALABLE PPRL

- filtering for specific similarity metrics / thresholds to reduce number of comparisons
  - privacy-preserving PPJoin (P4Join)
  - metric space: utilize triangular inequality
- (private) blocking approaches
  - partition datasets such that only records from same partition (block) need to be matched with each other
  - blocking at data owner on unencoded data (e.g., soundex) or at LU on bloom filters (e.g., LSH)
- parallel linkage
  - GPU-based matching of bit vectors
  - parallel matching on Hadoop clusters

EVALUATION RESULTS

- Configurations
  - two input datasets R, S determined with FEBRL data generator N=[100.000, 200.000, ..., 500.000]. |R|=1/5·N, |S|=4/5·N
  - bit vector length: 1000
  - similarity threshold 0.8
- runtime in minutes on standard PC

Approach	Dataset size N				
	100 000	200 000	300 000	400 000	500 000
NestedLoop	6.1	27.7	66.1	122.0	194.8
MultiBitTree	4.7	19.0	40.6	78.2	119.7
P4Join	2.3	15.5	40.1	77.8	125.5

- similar results for P4Join and Multibit Tree
- relatively small improvements compared to NestedLoop

Z. Sehili, L. Kolb, C. Borgs, R. Schnell, E. Rahm: *Privacy Preserving Record Linkage with PPJoin.* Proc. BTW Conf. 2015

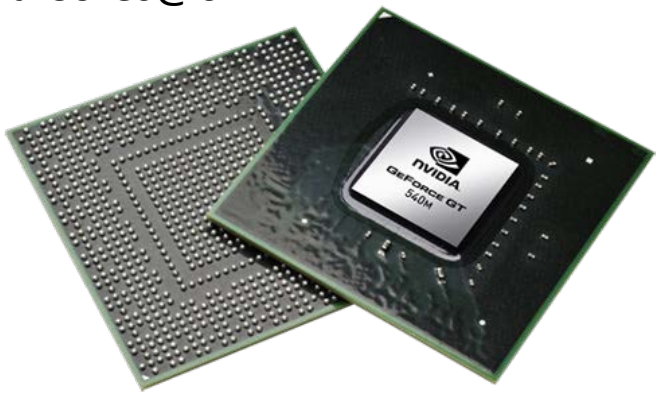
GPU-BASED RESULTS

GeForce GT 610

- 48 Cuda Cores@810MHz
- 1GB
- 35€

GeForce GT 540M

- 96 Cuda Cores@672MHz
- 1GB

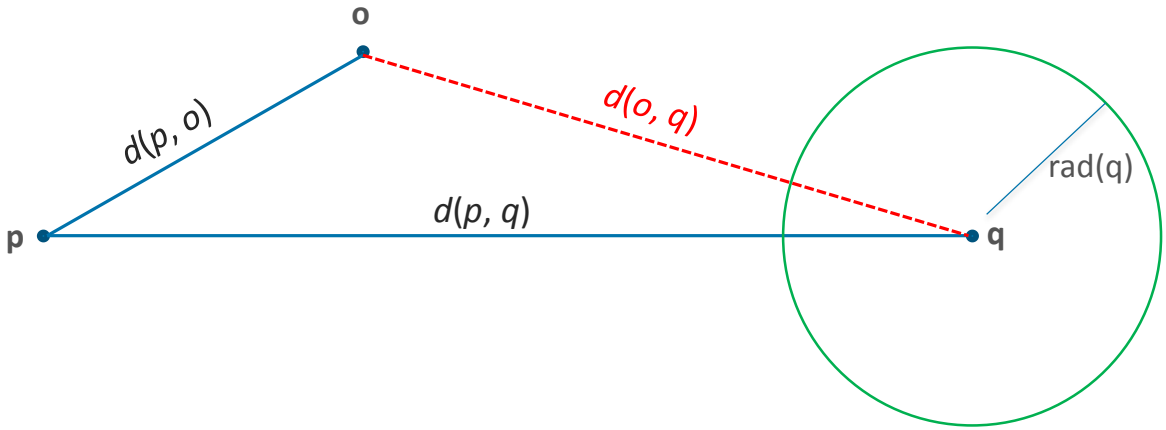


	100 000	200 000	300 000	400 000	500 000
GForce GT 610	0.33	1.32	2.95	5.23	8.15
GeForce GT 540M	0.28	1.08	2.41	4.28	6.67

- improvements by up to a factor of 20, despite low-profile graphic cards

METRIC SPACE-BASED DISTANCE FUNCTIONS

- distance functions d for metric spaces (e.g. Edit or Hamming distance, Jaccard coefficient) obey the triangular inequality



- can be used to reduce number of comparisons to find all entities within a maximal distance, e.g., using a *pivot-based* approach
  - select certain number of pivot objects in source D1 and assign each object to closest pivot
  - for each object from D2 only a subset of pivots and for each pivot only a subset of the assigned objects need to be considered for finding matches

Z. Sehili, E. Rahm: *Speeding Up Privacy Preserving Record Linkage for Metric Space Similarity Measures.* Datenbankspektrum 2016

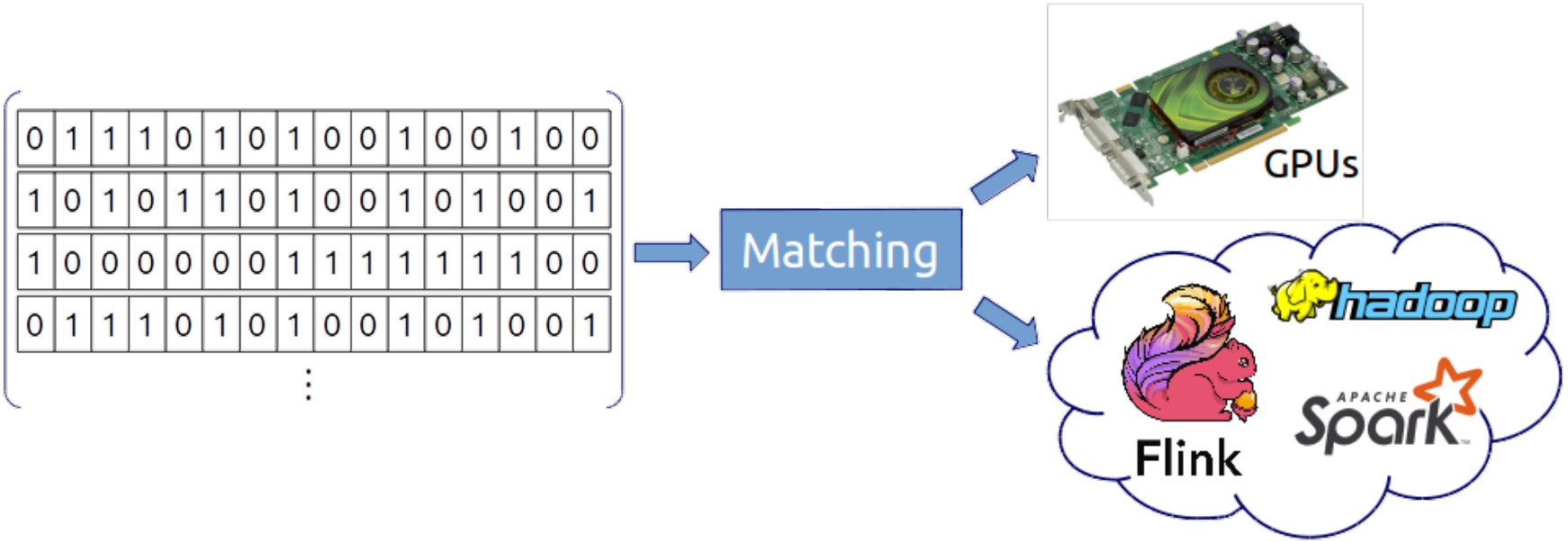
EVALUATION

- comparison with previous approaches using the same datasets
- runtime in minutes (using faster PC than in previous evaluation)

Algorithms	Datasets				
	100 000	200 000	300 000	400 000	500 000
NestedLoop	3.8	20.8	52.1	96.8	152.6
MultiBitTree	2.6	11.3	26.5	50.0	75.9
P4Join	1.4	7.4	24.1	52.3	87.9
Pivots (metric space)	0.2	0.4	0.9	1.3	1.7

- pivot-based approach performs best with up to 40X faster than other algorithms
- still quadratic increase with #records
  - run time for 16 million records ?

PARALLEL PPRL

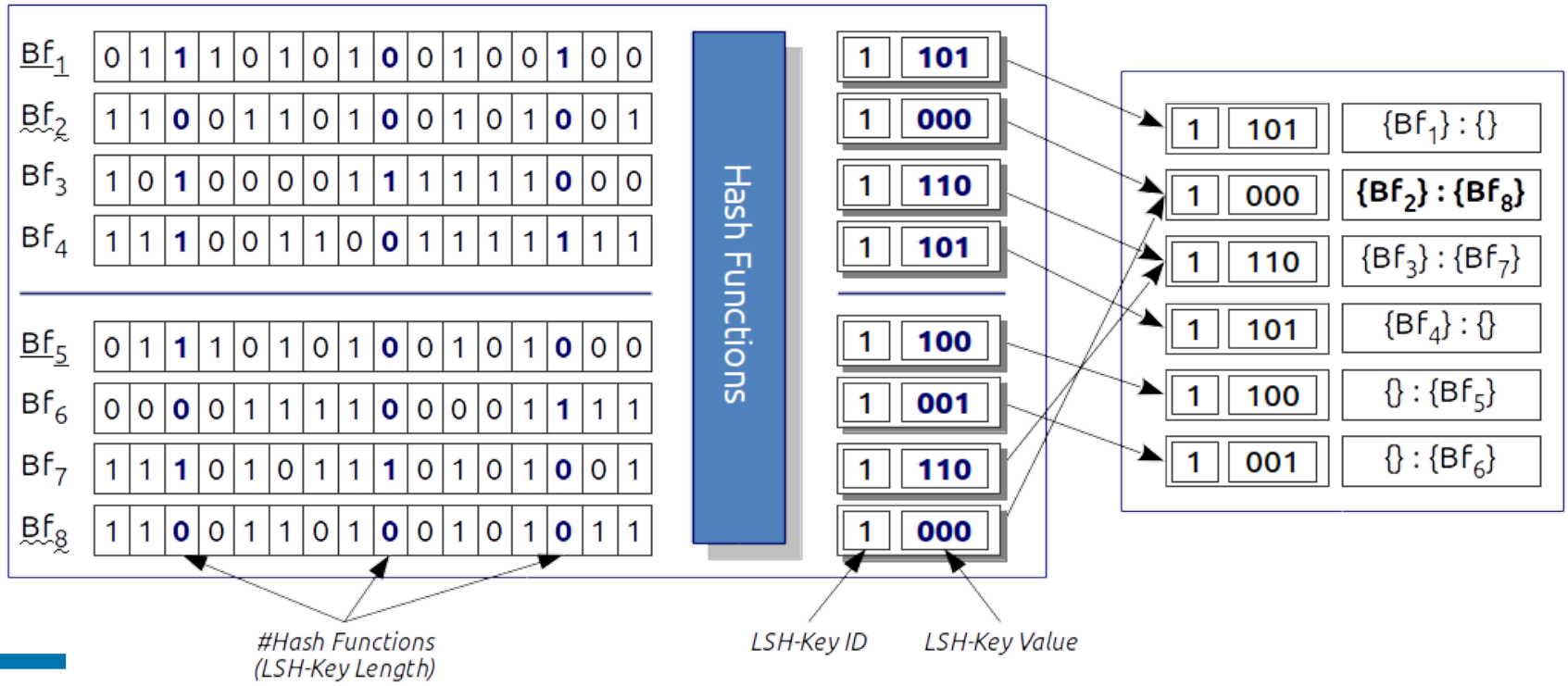


- Speed up PPRL by
  - using GPUs or/and
  - distributed processing frameworks
    (+ blocking/filtering)

Locality Sensitive Hashing (LSH)

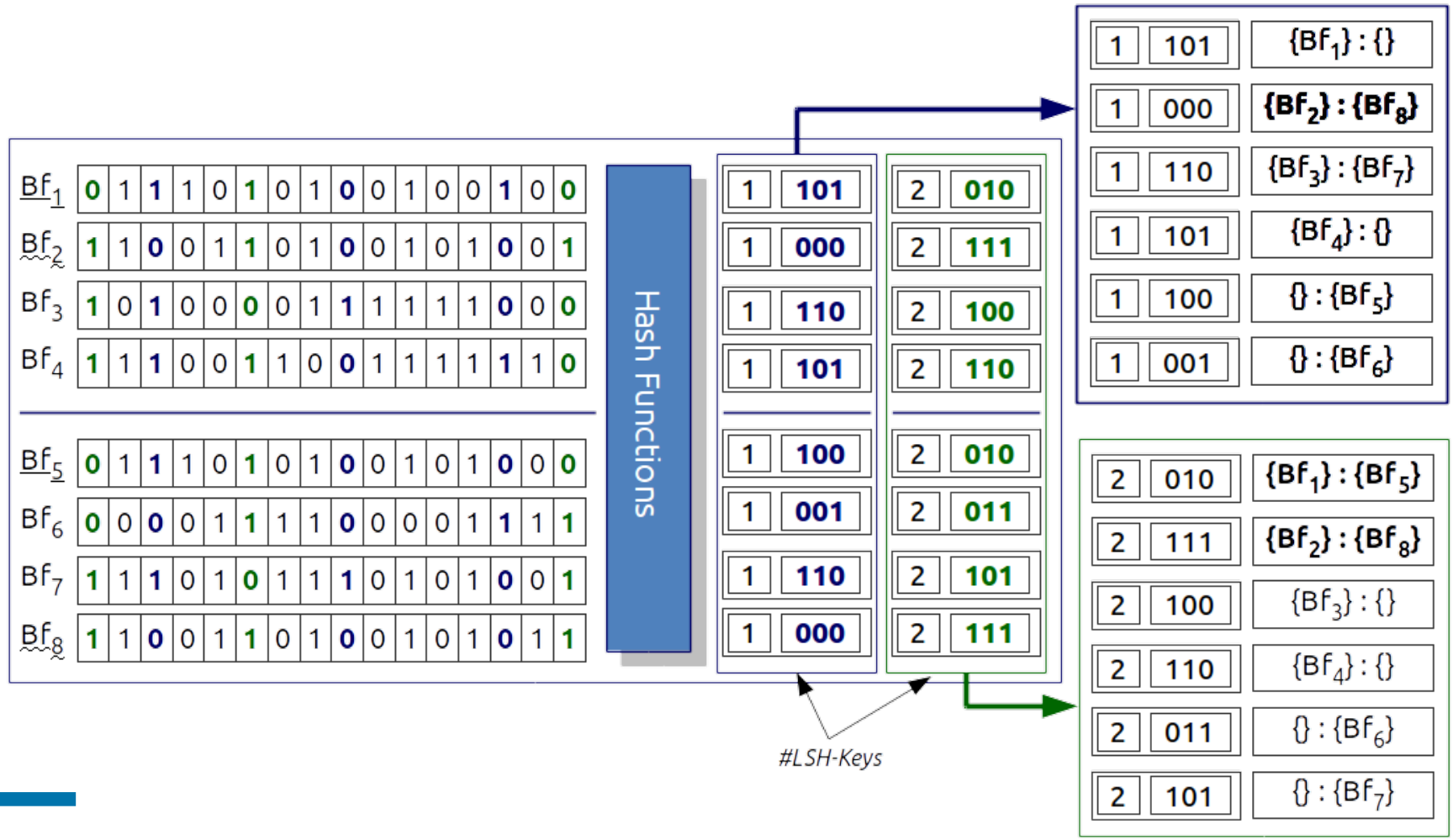
- Probabilistic blocking using *k* locality sensitive hash functions (works for Jaccard similarity, Hamming distance)
- Concatenation of *k* hash values as blocking key

Example for k=3, matches: (Bf_1, Bf_5), (Bf_2, Bf_8)

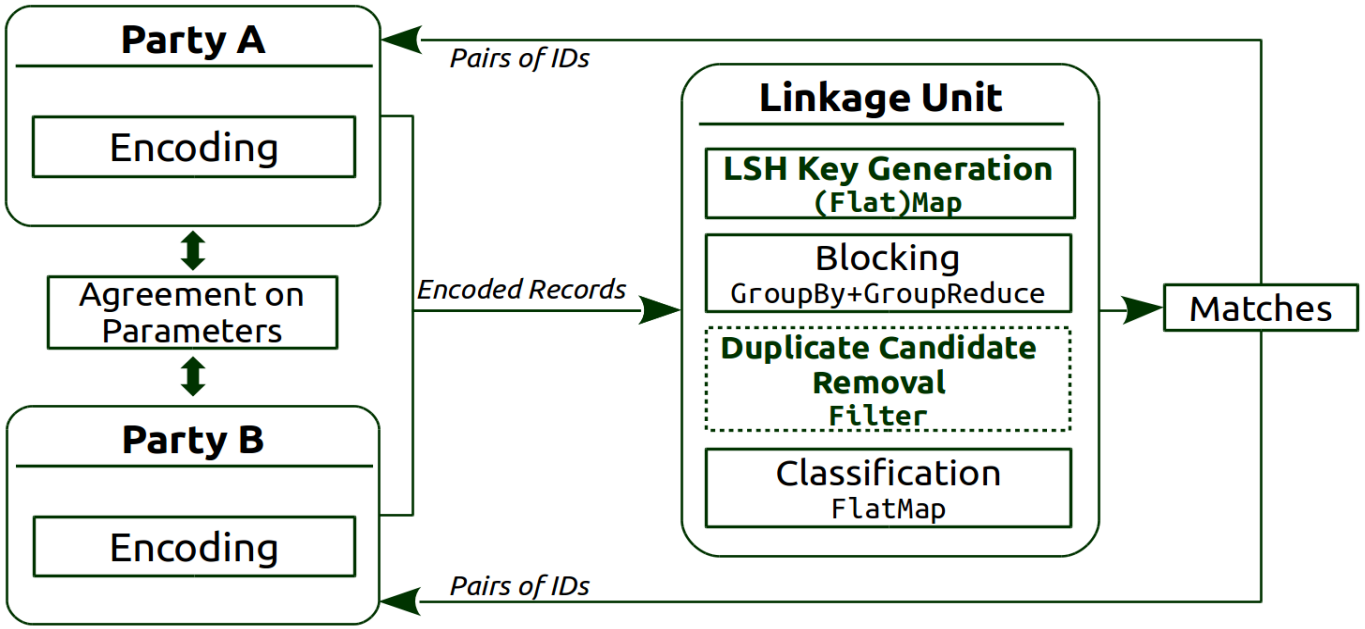


LSH (2)

- Due to dirty data multiple (*m*) LSH-Keys are necessary



PPRL WORKFLOW USING APACHE FLINK

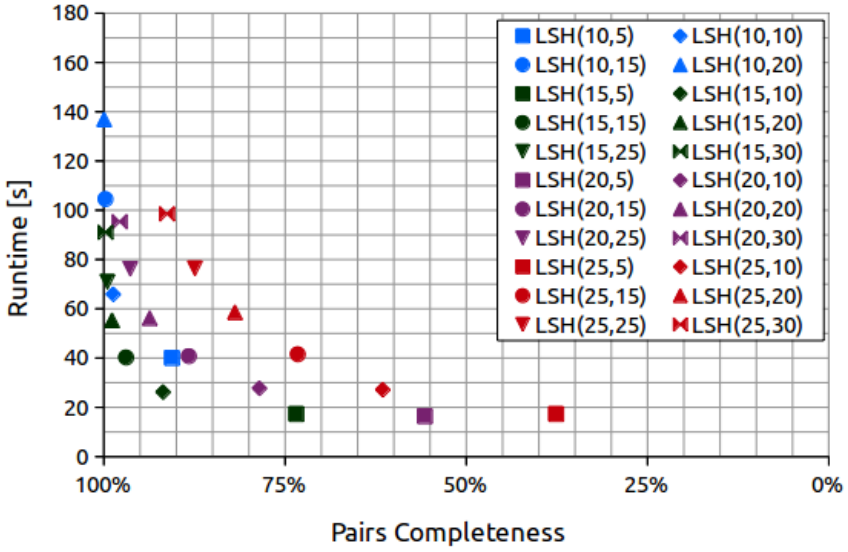


- generation of LSH blocking key (BK) at the LU
  - optional elimination of duplicate candidates for multiple BKs

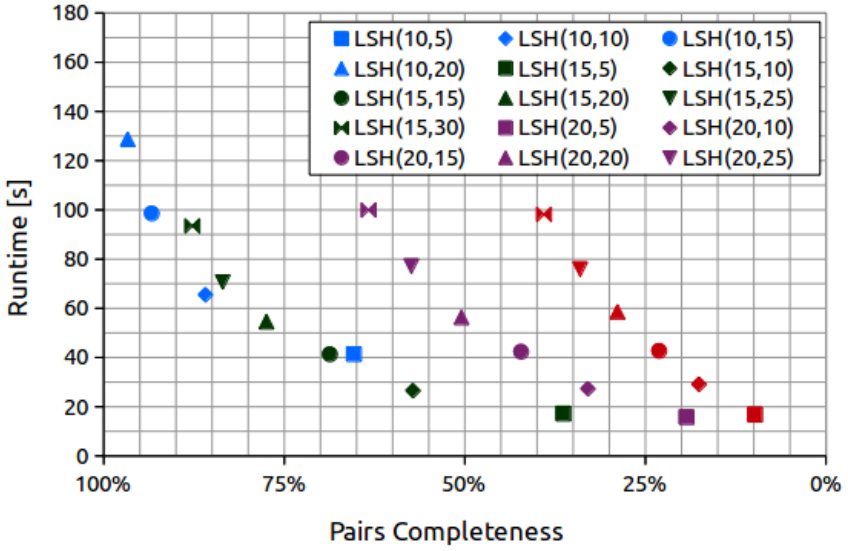
Evaluation

- generated data sets with different corruption levels
  - *Moderate*: 2 modifications (max. 1 per attribute)
  - *Heavy*: up to 6 modifications (max. 2 per attribute)
- 1 million records
- encoded attributes: name, surname, date of birth, zip code and city

Moderate corruption level



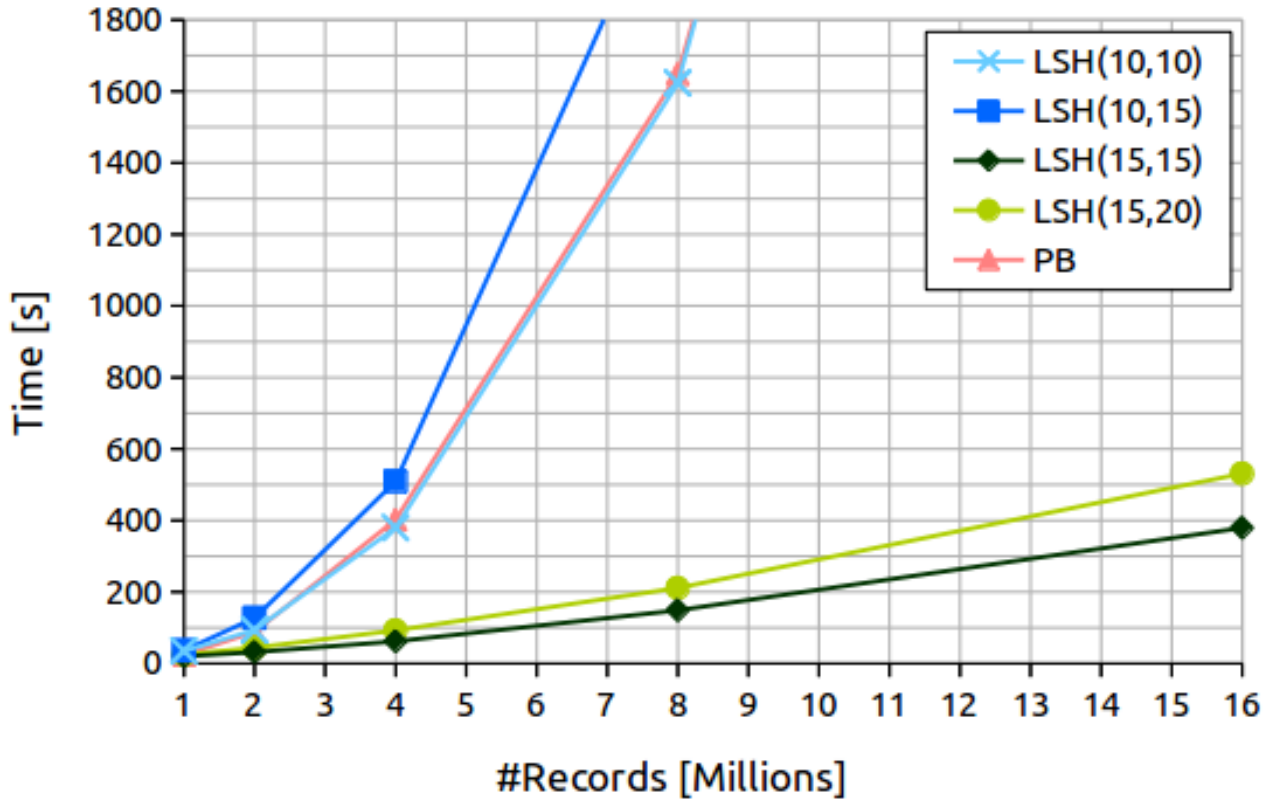
Heavy corruption level



LSH (k,m)

Scalability

- LSH with a key length of 15 clearly outperforms phonetic blocking

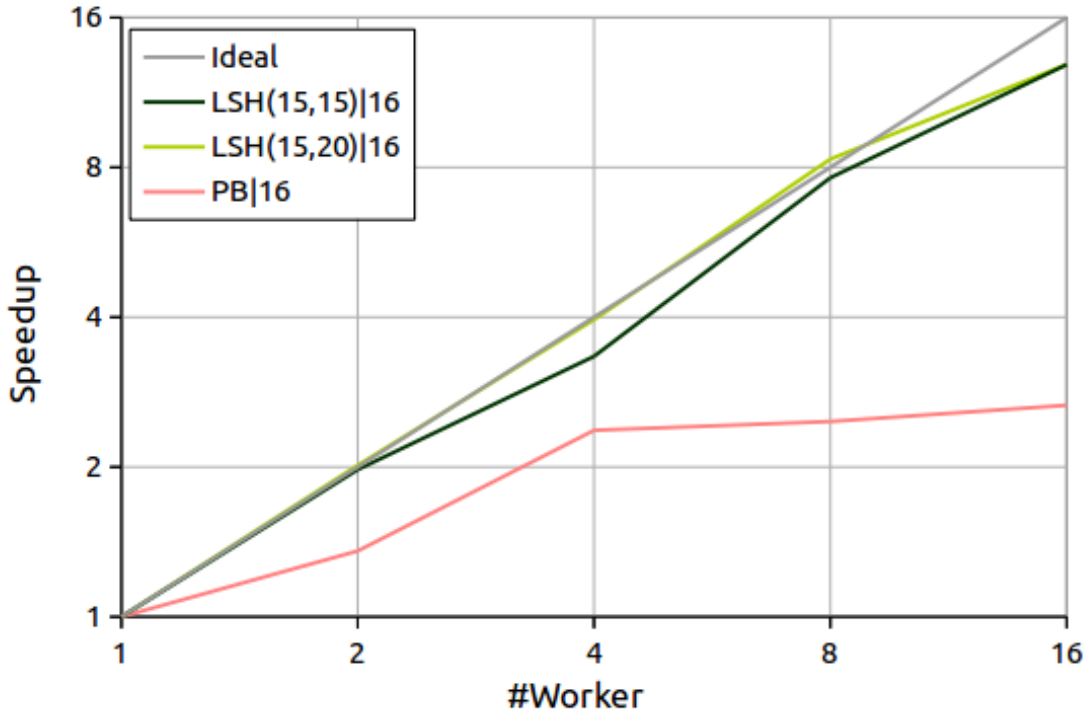


ScaDS DRESDEN LEIPZIG

Speedup

UNIVERSITÄT LEIPZIG

- utilizing up to 16 worker nodes
- nearly ideal speedup for LSH and up to 8 workers
- skew effects limit the possible speedup for phonetic blocking (large blocks for common names)



SUMMARY

- Privacy for Big Data
  - privacy-preserving publishing / record linkage / data mining
  - tradeoff between protection of personal/sensitive data and data utility for analysis
  - complete anonymization prevents record linkage -> 1-way pseudonymization of sensitive attributes good compromise
- Scalable Privacy-Preserving Record Linkage
  - bloom filters allow simple, effective and relatively efficient match approach
  - performance improvements by blocking / filtering / parallel PPRL
  - effective filtering by utilizing metric-space distance functions
  - GPU and cluster usage achieve significant speedups

OUTLOOK / CHALLENGES

- high PPRL match quality for real, dirty data
- quantitative evaluation of privacy characteristics
- efficient PPRL approaches for multiple sources with and without linkage unit
- combined study of PPRL + data mining
- more practical use cases

REFERENCES

- R. Agrawal et al: Hippocratic Databases. Proc. VLDB 2002
- B. Fung, K. Wang, R.Chen, P.S Yu: *Privacy-preserving data publishing: A survey of recent developments*. ACM CSUR 2010
- R. Hall, S. E. Fienberg: *Privacy-Preserving Record Linkage.* Privacy in Statistical Databases 2010: 269-283
- M. Franke, Z. Sehili, E. Rahm: *Parallel Privacy Preserving Record Linkage using LSH-based blocking*. Univ. Leipzig, 2017 (submitted)
- D. Karapiperis, V. S. Verykios: *A distributed near-optimal LSH-based framework for privacy-preserving record linkage.* Comput. Sci. Inf. Syst. 11(2): 745-763 (2014)
- C.W. Kelman, A.J. Bass. C.D. Holman: *Research use of linked health data--a best practice protocol*. Aust NZ J Public Health. 2002
- R. Schnell, T. Bachteler, J. Reiher: *Privacy-preserving record linkage using Bloom filters*. BMC Med. Inf. & Decision Making 9: 41 (2009)
- Z. Sehili, L. Kolb, C. Borgs, R. Schnell, E. Rahm: *Privacy Preserving Record Linkage with PPJoin*. Proc. BTW Conf. 2015
- Z. Sehili, E. Rahm: *Speeding Up Privacy Preserving Record Linkage for Metric Space Similarity Measures*. Datenbankspektrum 2016
- D. Vatasalan, P. Christen, C.M. O'Keefe, V.S. Verykios: *A taxonomy of privacy-preserving record linkage techniques*. Information Systems 2013
- D. Vatasalan, P. Christen, E. Rahm: *Scalable privacy-preserving linking of multiple databases using Counting Bloom filters*. Proc Privacy and Discrimination in Data Mining (PDDM), 2016
- D. Vatasalan, Z. Sehili, P- Christen, E. Rahm: *Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges*. In: Handbook of Big Data Technologies, Springer 2017