

Lu, Yang ORCID logoORCID:

<https://orcid.org/0000-0002-0583-2688> (2022) Privacy-preserving access control in electronic health record linkage. In: International Population Data Linkage Network, September 7-9, 2022, Edinburgh, UK.

Downloaded from: <https://ray.yorks.ac.uk/id/eprint/7663/>

Research at York St John (RaY) is an institutional repository. It supports the principles of open access by making the research outputs of the University available in digital form. Copyright of the items stored in RaY reside with the authors and/or other copyright owners. Users may access full text items free of charge, and may download a copy for private study or non-commercial research. For further reuse terms, see licence terms governing individual outputs. [Institutional Repository Policy Statement](#)

RaY

Research at the University of York St John

For more information please contact RaY at ray@yorks.ac.uk

PRIVACY-PRESERVING ACCESS CONTROL IN ELECTRONIC HEALTH RECORD LINKAGE

Yang Lu

York St John University

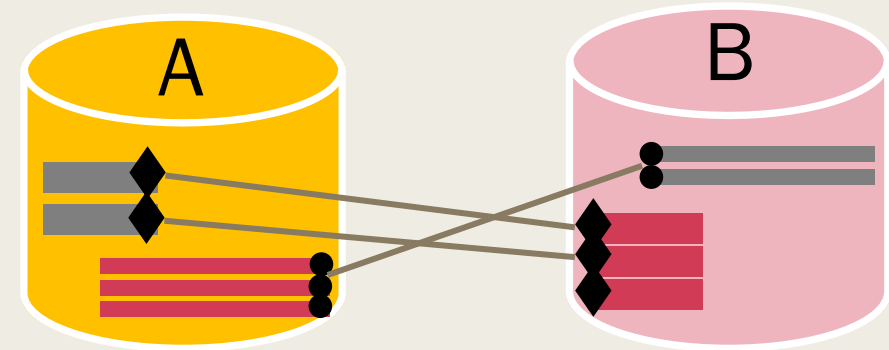
Outline

- Background
- Method
- Experiment
- Conclusion

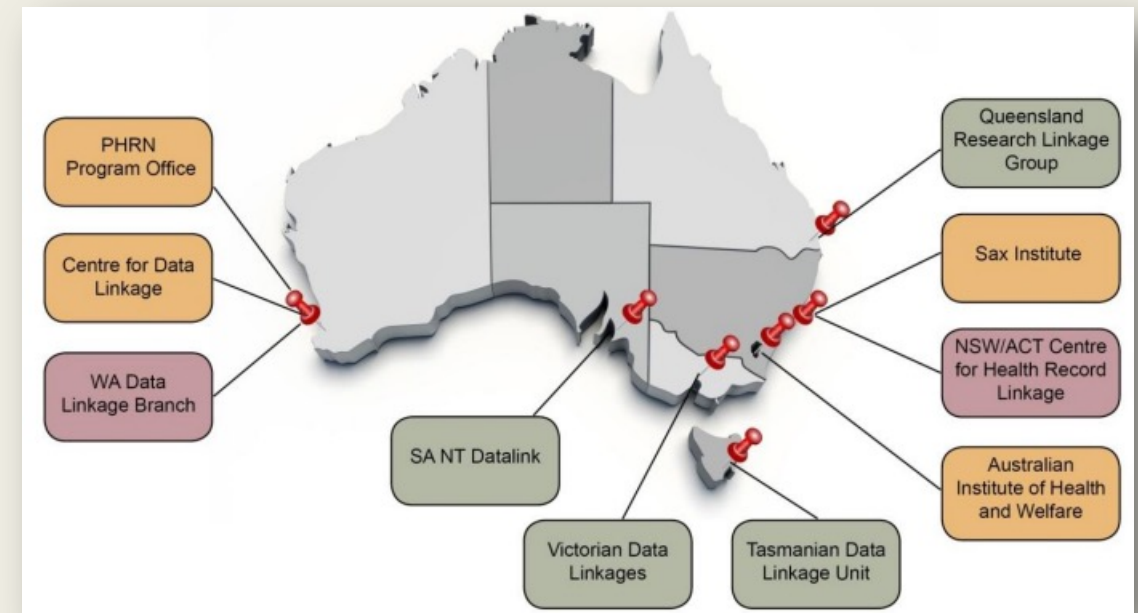
BACKGROUND

Health Record Linkage

Record linkage refers to the technique matching records about same entities across sources (e.g. data files, books, websites, and databases).



Data Linkage
WESTERN AUSTRALIA



Protected Personal Information



Eliminating 18 identifiers from personal health information to ensure no one can be identified



Patient identifiers are categorised into individually identifiable, re-identifiable and non-identifiable.



GDPR applies to any information relating to an identified or identifiable natural person

	HIPAA protected health information
1	Names
2	Zip Code
3	Dates MM/DD/YYYY
4	Phone numbers
5	Fax numbers
6	E-mail address
7	SSNs
8	MRN numbers
9	Insurance ID #s
10	Account #s
11	Certificate / License
12	Serial #s
13	Device #s
14	URL
15	IP address
16	Biometrics
17	Photos
18	Other

Statistical Disclosure Control

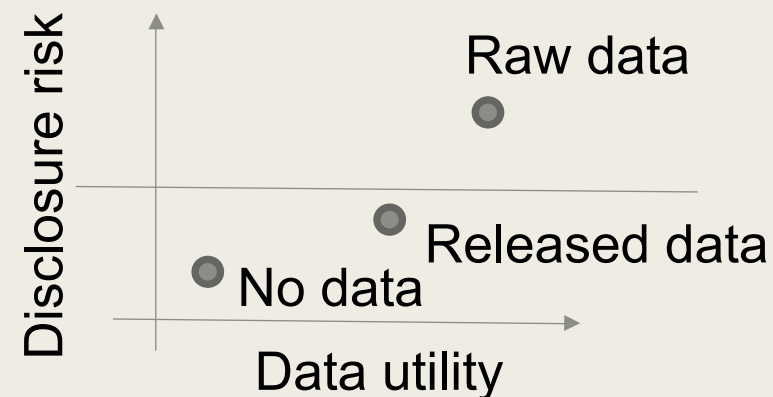
- “Individuals should not be uniquely identified”
- 87% of the US population could be re-identified by combining de-identified data sets (Sweeney, 2000)
- *Statistical Disclosure Control* (SDC) refers to a family of statistic-based technique are studied to ensure no individual can be re-identified.
 - *K-anonymity and its variants*
 - *Differential privacy*

Challenges

- Delay
 - *Decentralisation?*
- Opaque
 - *Transparent, verifiable, quantifiable?*
- Priori knowledge
 - *(non-malicious) Disclosure risk detection*
 - *Minimizing utility loss*

Privacy Preservation with *K-Anonymity*

R-U confidential map:



Quasi-identifiers (QIs)

	Zip code	Age	Nationality
1	13053	28	Russian
2	13068	29	American
3	13068	21	Japanese
4	13053	23	American
5	14853	50	Indian
6	14853	55	American
7	14850	47	American
8	14850	49	American

4-anonymity

	Zip code	Age	Nationality
1	130**	<30	*
2	130**	<30	*
3	130**	<30	*
4	130**	<30	*
5	1485*	≥40	*
6	1485*	≥40	*
7	1485*	≥40	*
8	1485*	≥40	*

Raw records

	Zip code	Age	Nationality
1	13053	28	Russian
2	13068	29	American
3	13068	21	Japanese
4	13053	23	American
5	14853	50	Indian
6	14853	55	American
7	14850	47	American
8	14850	49	American

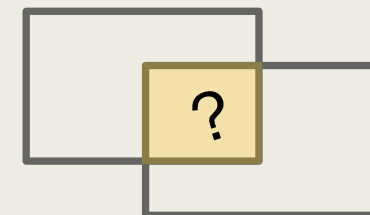
Released records

	Zip code	Age	Nationality
1	130**	<30	*
2	130**	<30	*
3	130**	<30	*
4	130**	<30	*
5	1485*	≥40	*
6	1485*	≥40	*
7	1485*	≥40	*
8	1485*	≥40	*

4-anonymity

The “presence” of individuals is **known** in k-anonymity case

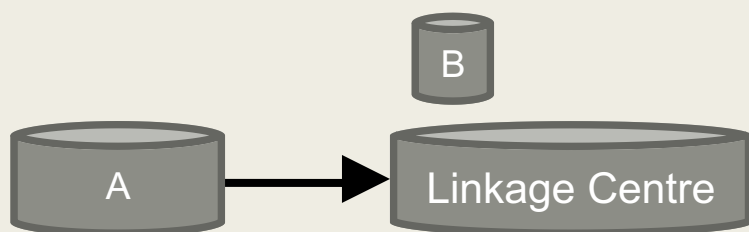
Linkage disclosure - “presence” is unknown



Weak k-anonymity (Atzori, 2006)

“having the same goal as k-anonymity: released tuples need to match at least k individuals back to original dataset.”

Linkage 3-Anonymity (*Time n*)



(Linkage) QIs Non-QI

	YoB	Sex	Ethnicity	Language
1	1974	F	Chinese	Mandarin
2	1980	F	Chinese	Mandarin
4	1968	M	Malaysian	English

	YoB	Sex	Ethnicity
1	1970-1980	F	Chinese
2	1960-1980	F	Chinese
3	1960-1980	*	Malaysian
4	1970-1980	*	Malaysian
5	1970-1980	F	Chinese
6	1960-1980	*	Malaysian

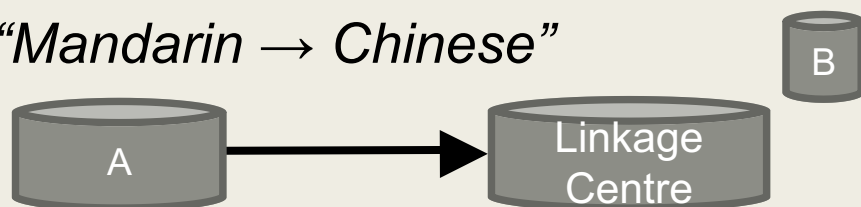
	YoB	Sex	Ethnicity	Language
1	1970-1980	F	Chinese	Mandarin
2	1960-1980	F	Chinese	Mandarin
3	1960-1980	*	Malaysian	English

"Mandarin → Chinese"



Linkage 3-Anonymity (*Time $n+1$*)

"Mandarin \rightarrow Chinese"



(Linkage) QIs

	YoB	Sex	Ethnicity	Language
1	1974	F	Chinese	Mandarin
2	1980	F	Chinese	Mandarin
4	1968	M	Malaysian	English

	YoB	Sex	Ethnicity	Language
1	1960-1980	*	Asian	Mandarin
2	1960-1980	*	Asian	Mandarin
3	1960-1980	*	Asian	Mandarin
4	1960-1980	*	Asian	English
5	1960-1980	*	Asian	English
6	1960-1980	*	Asian	English

	YoB	Sex	Ethnicity	Language
1	1960-1980	*	Chinese	Mandarin
2	1960-1980	*	Chinese	Mandarin
4	1960-1980	*	Asian	English

Design Process - SLKA

Est.
1841

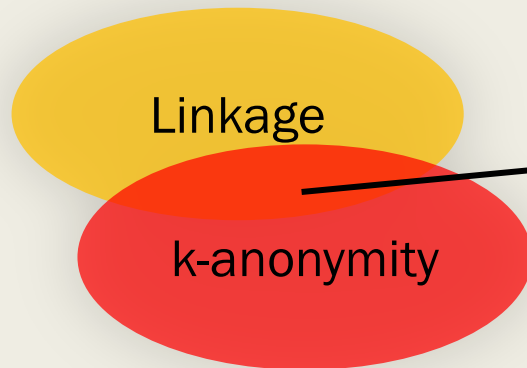
YORK
ST JOHN
UNIVERSITY

Baseline k-anonymity



Anonymity Scheme = $\{QI, k\}$

Linkage k-anonymity

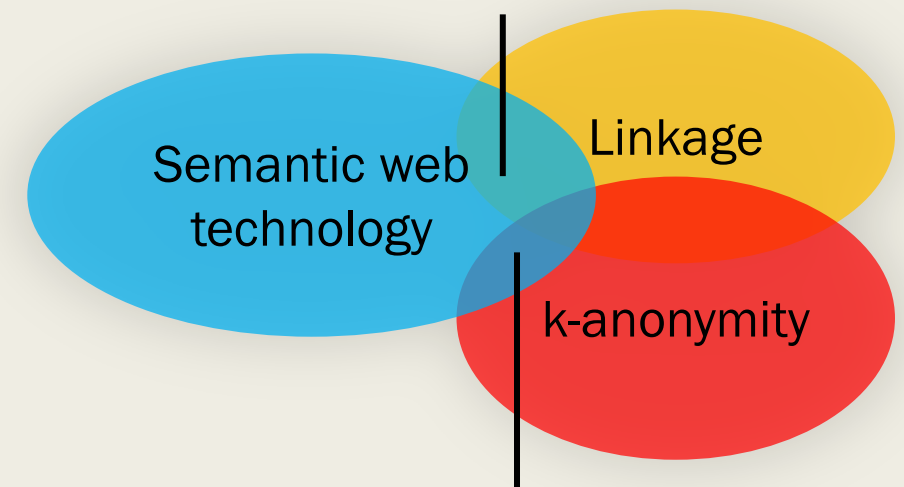


Less information
loss

Anonymity Scheme = $\{LQI, k_{max}\}$

Semantic-based linkage k-anonymity (SLKA)

Privacy verification
in dynamic datasets



Policy/Rule composition
to arbitrary linkage

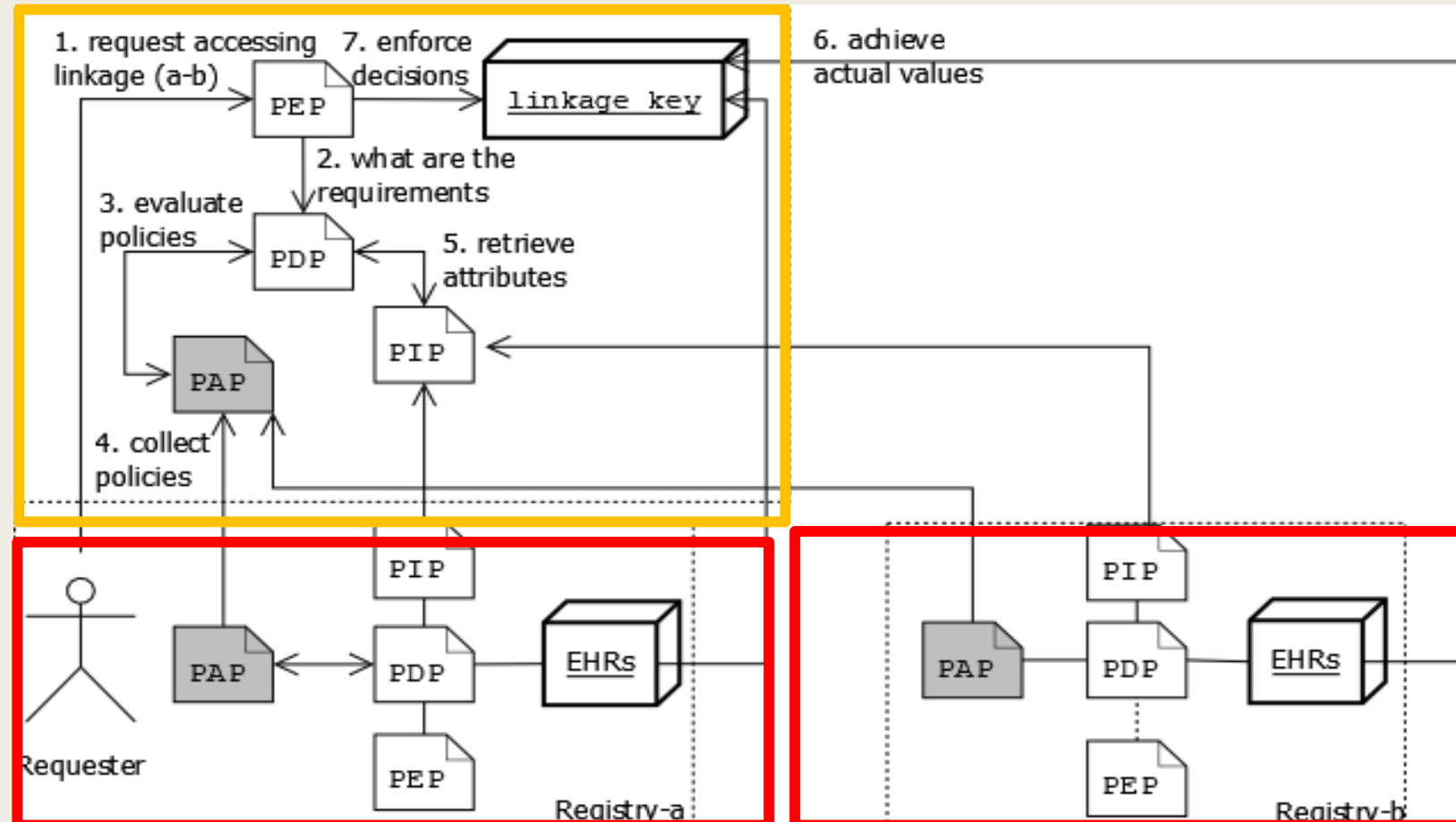
Anonymity Scheme = $\{LQI, k_{max}, Associations_{12}\}$

Est.
1841

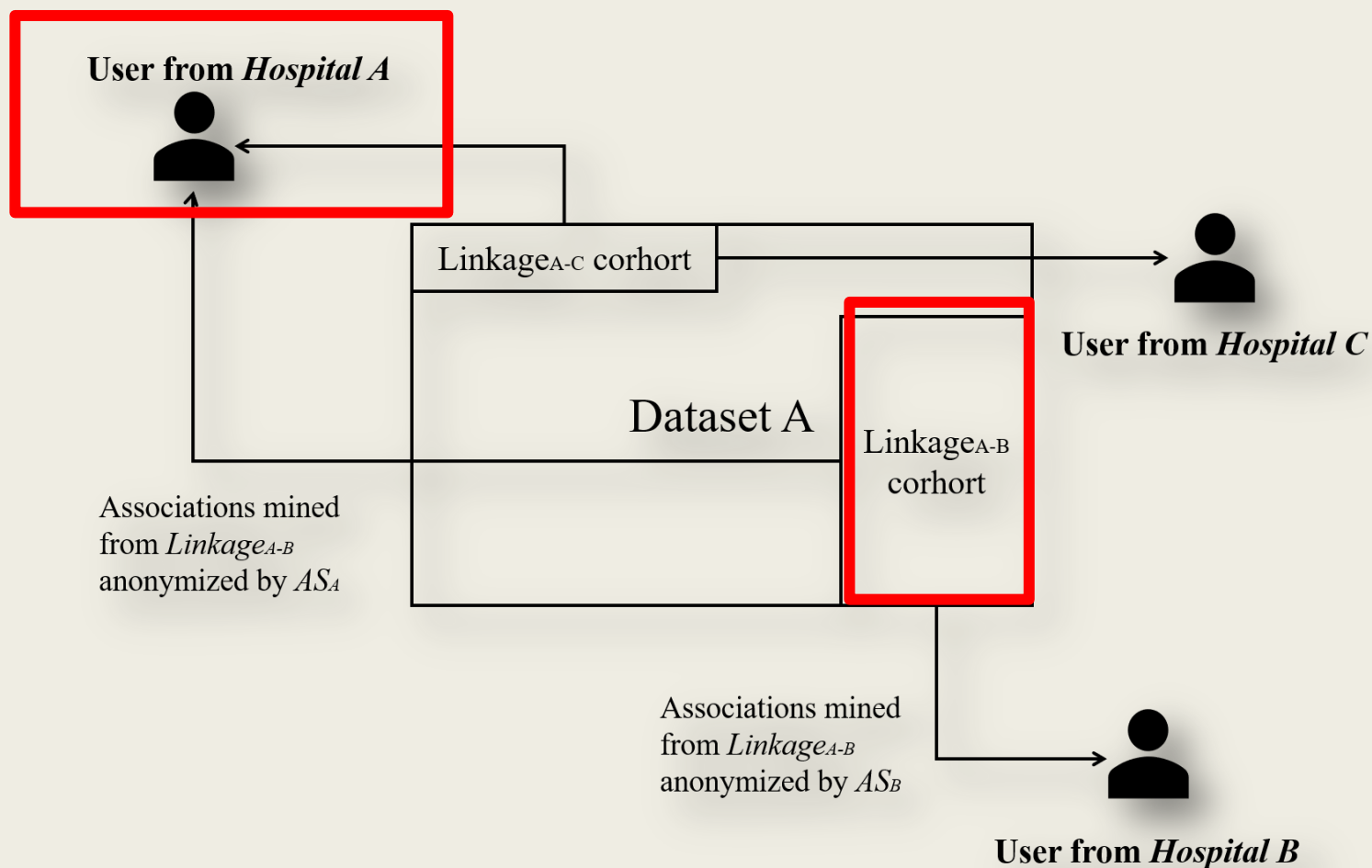
YORK
ST JOHN
UNIVERSITY

METHOD

Extending XACML Framework

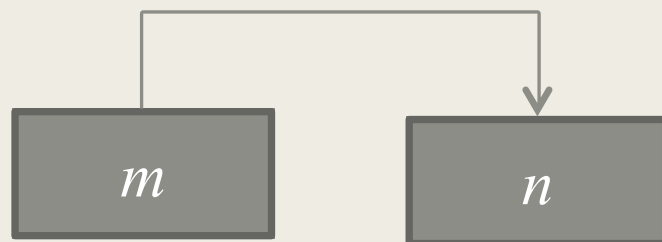


Role-based Knowledge Management

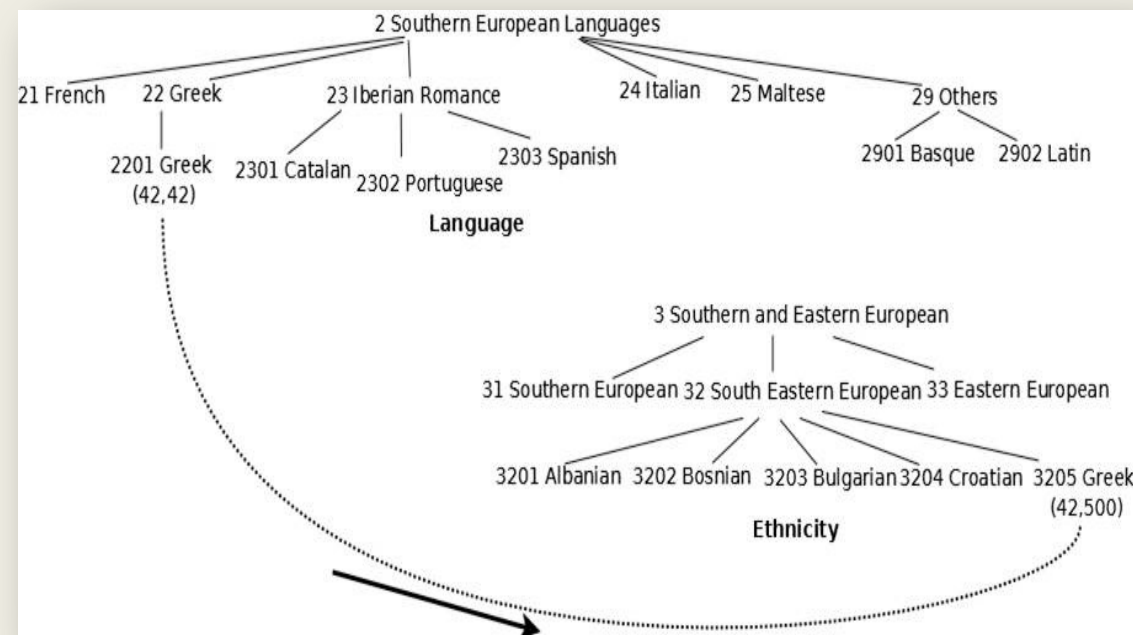


Semantic-based Inference Control

- Mining associations - Apriori
- Minimum Support (ms) and Confidence (mc)
- Conditionals – NQI (m); Consequence – LQI (n);



$$C_1^m \cdot C_1^n$$



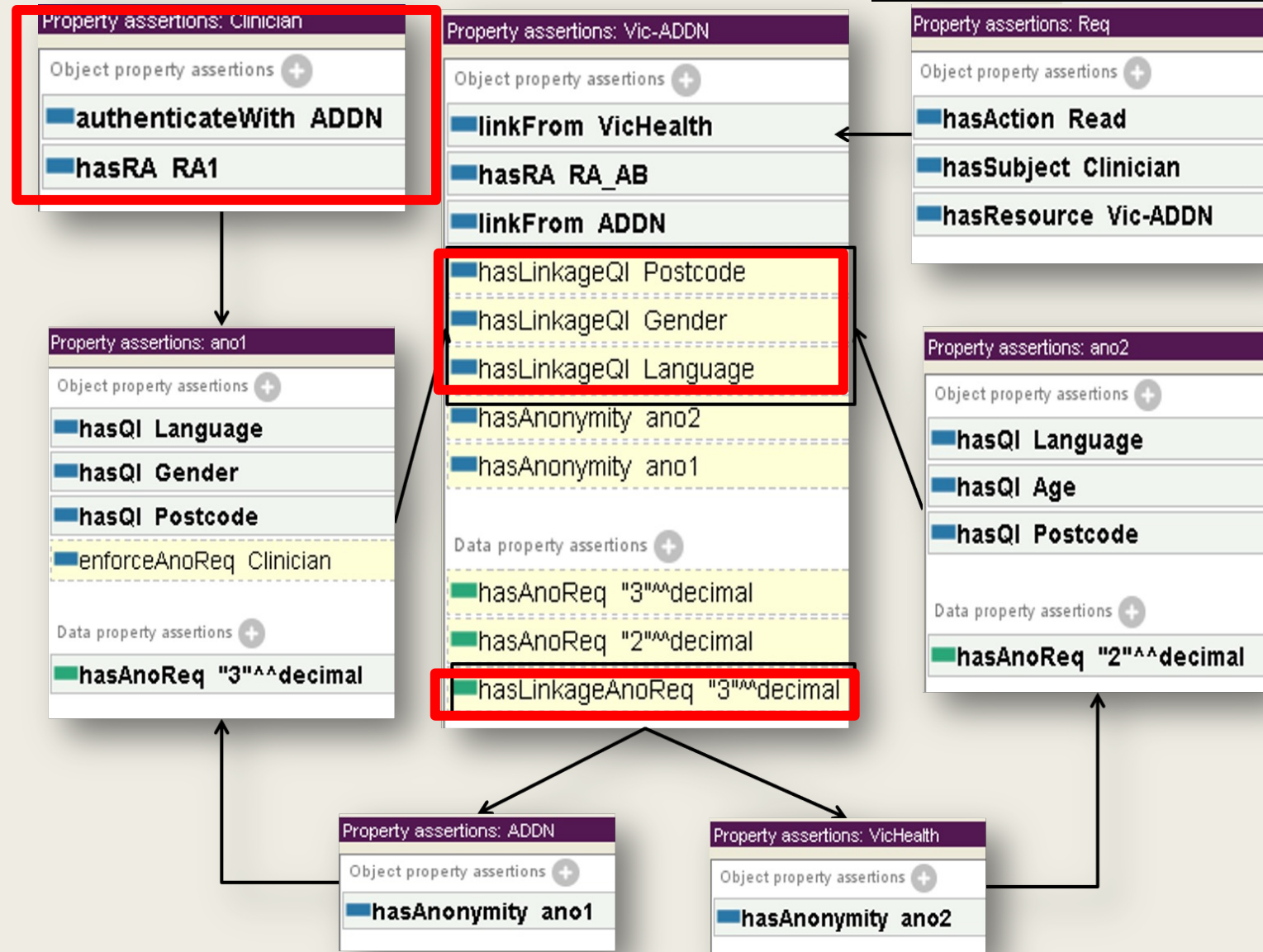
Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

Formalisation and Evaluation (1/2)

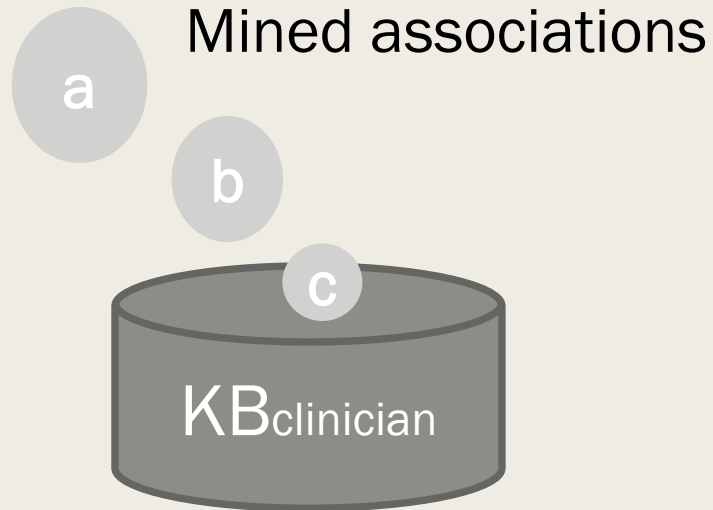
Est.
1841

YORK
ST JOHN
UNIVERSITY

Requestor (role) → Policy scheming
→ Privacy verification



Formalisation and Evaluation (2/2)



NQI -> *LQI*

e.g. “2201 Greek” -> “3205 Greek”

Property assertions: tuple1	
Object property assertions +	
hasAttribute	2201-Greek
hasAttribute	32-Southern_Eastern_European

Property assertions: Generalisation	
Object property assertions +	
enforce	22-Greek
enforce	2201-Greek

Property assertions: tuple2	
Object property assertions +	
hasAttribute	22-Greek
hasAttribute	32-Southern_Eastern_European

Property assertions: Obligation1	
Object property assertions +	
hasTarget	Tar1
enforceRA	RA1
enforceAnonymity	an1

Triggering RA1

Property assertions: RA1	
Object property assertions +	
hasFunction	Generalisation
hasConsequenceAttr	3205_Greek
isA	RA_AB
hasSubRA	SubRA1
hasConditionAttr	2201_Greek

Triggering SubRA1

Property assertions: SubRA1	
Object property assertions +	
hasConsequenceAttr	3205-Greek
hasFunction	Generalisation
hasConditionAttr	22-Greek

Available tuple1

RA1+tuple1

SubRA1+tuple2

Temporary tuple2

Est.
1841

YORK
ST JOHN
UNIVERSITY

EXPERIMENT

Simulation

Est.
1841

YORK
ST JOHN
UNIVERSITY

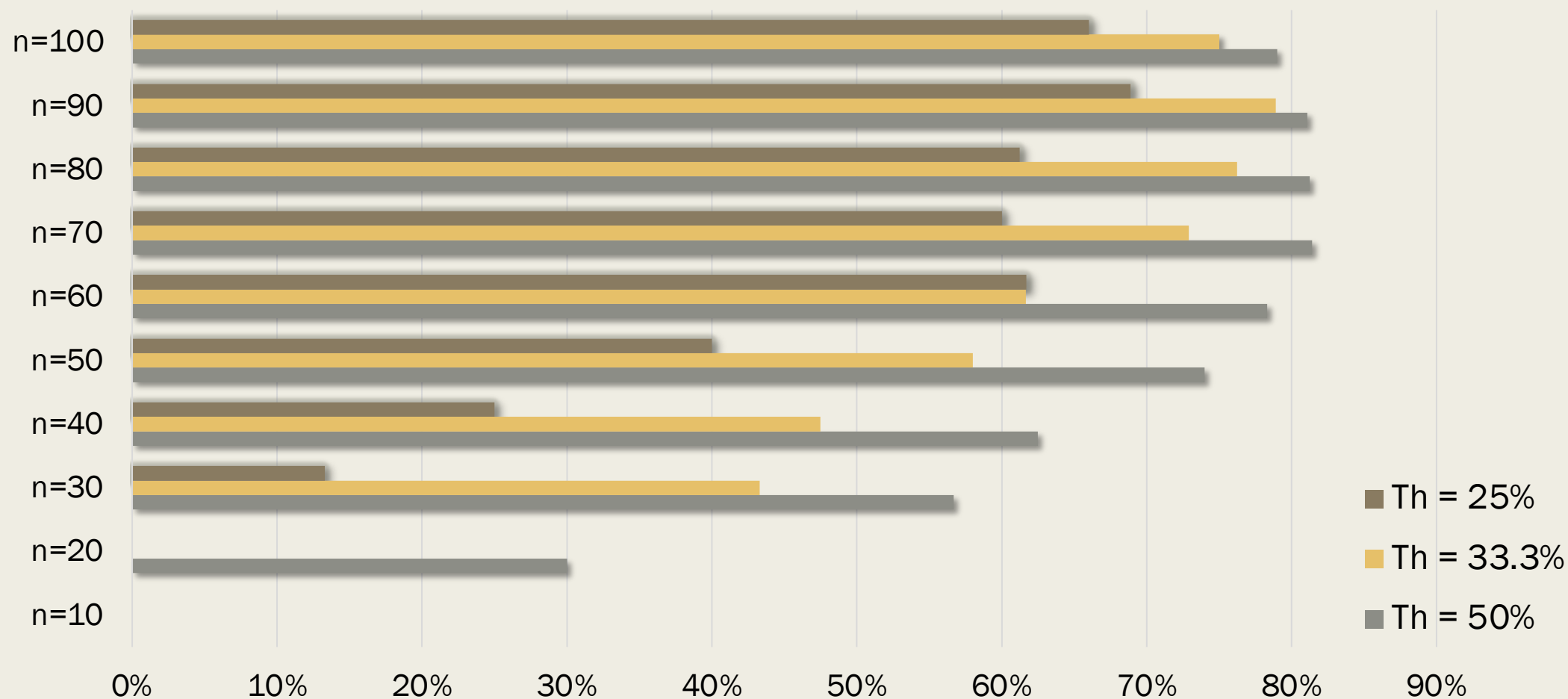
- 1000 (996 after cleaning) synthetic records about VicHealth survey respondents
- Label records as the “linked” (10 linkage datasets)
- Simulating *repeated linkage* requests (Time1 & Time 2)
 - Same: VicHealth user (role), candidate datasets, cohort, identifiers
 - Different: policies (Anonymity schemes)
 - Time1 {age, ethnicity, postcode} and Time2 {age, ethnicity, postcode, language}
- $k = 2, 3, 4$

1% - 10%

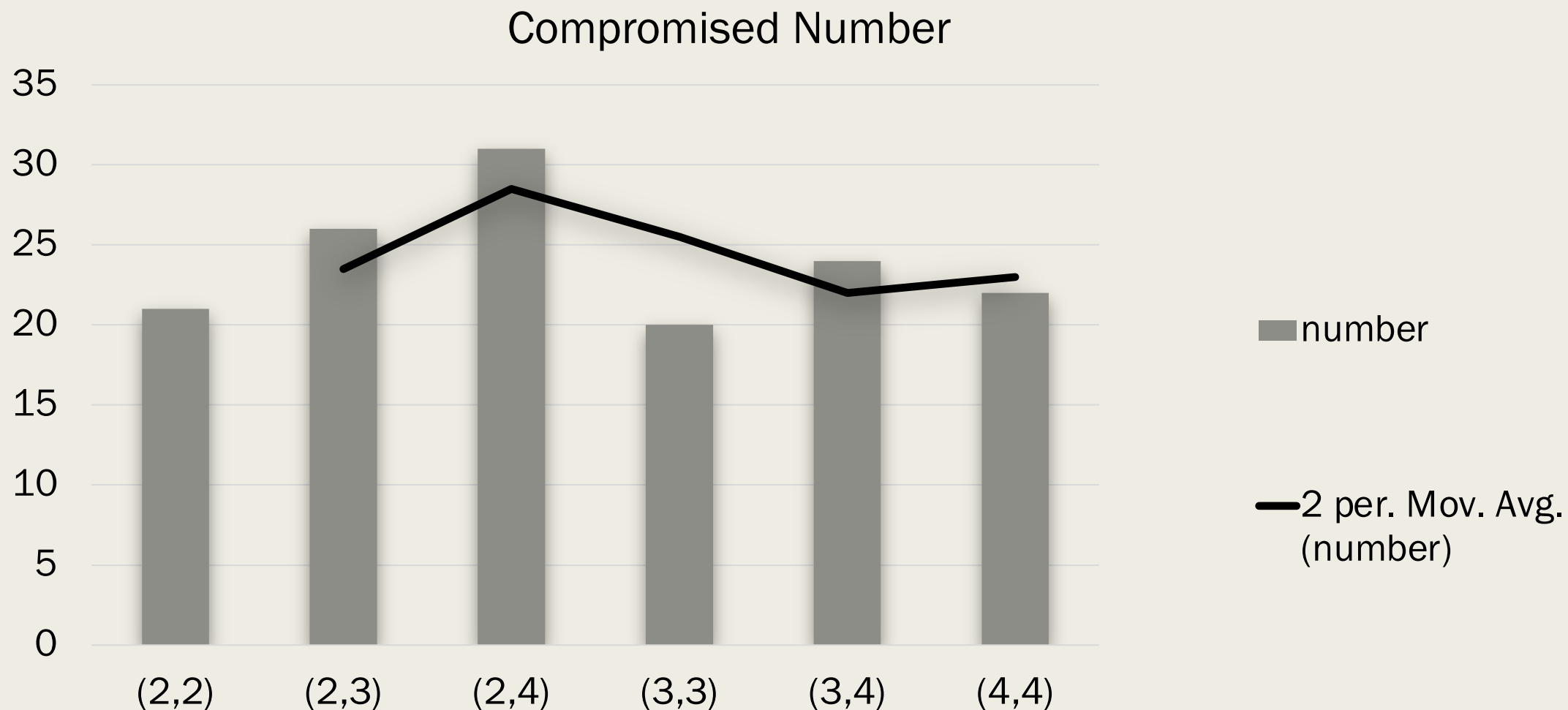
1000

Security Condition

Records satisfying the privacy requirement

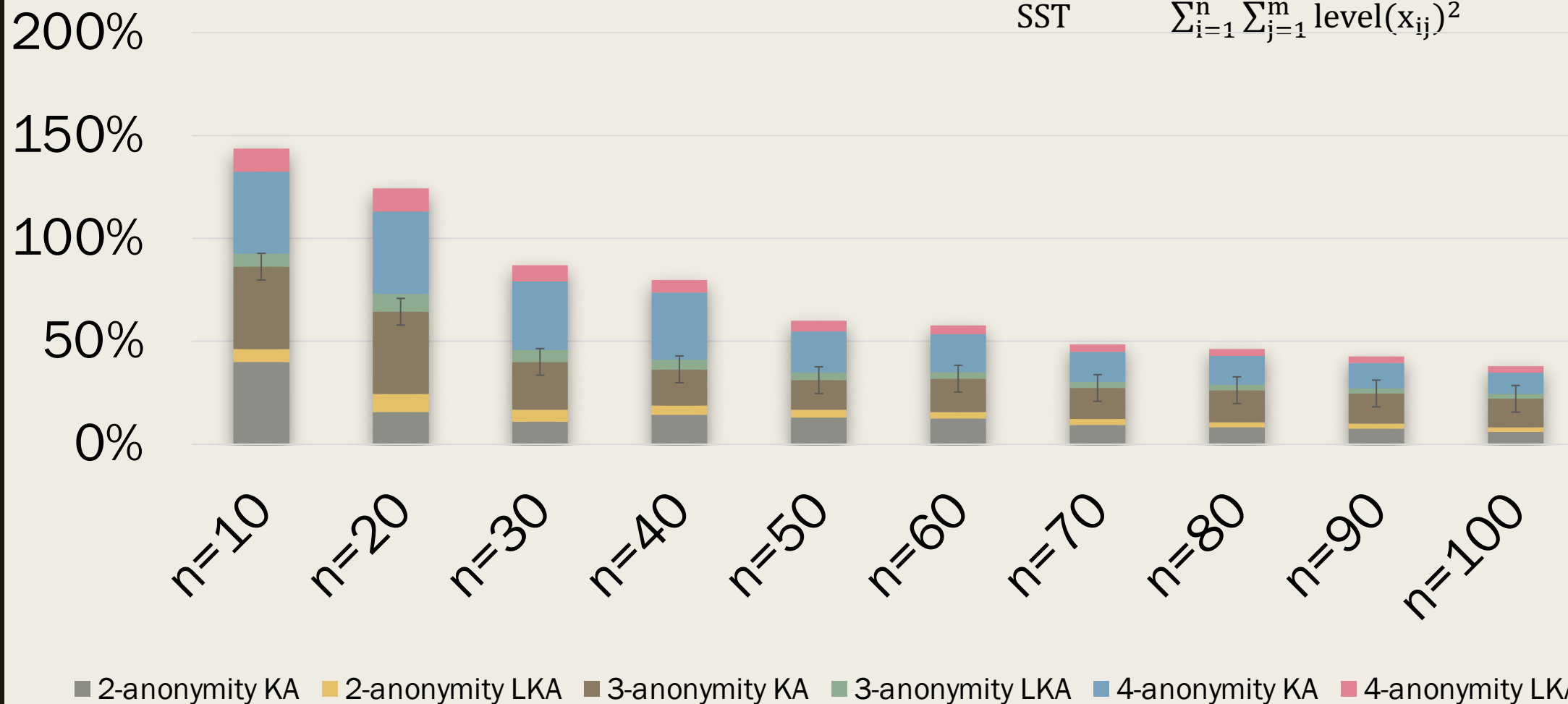


Privacy Violation – Weak k-anonymity

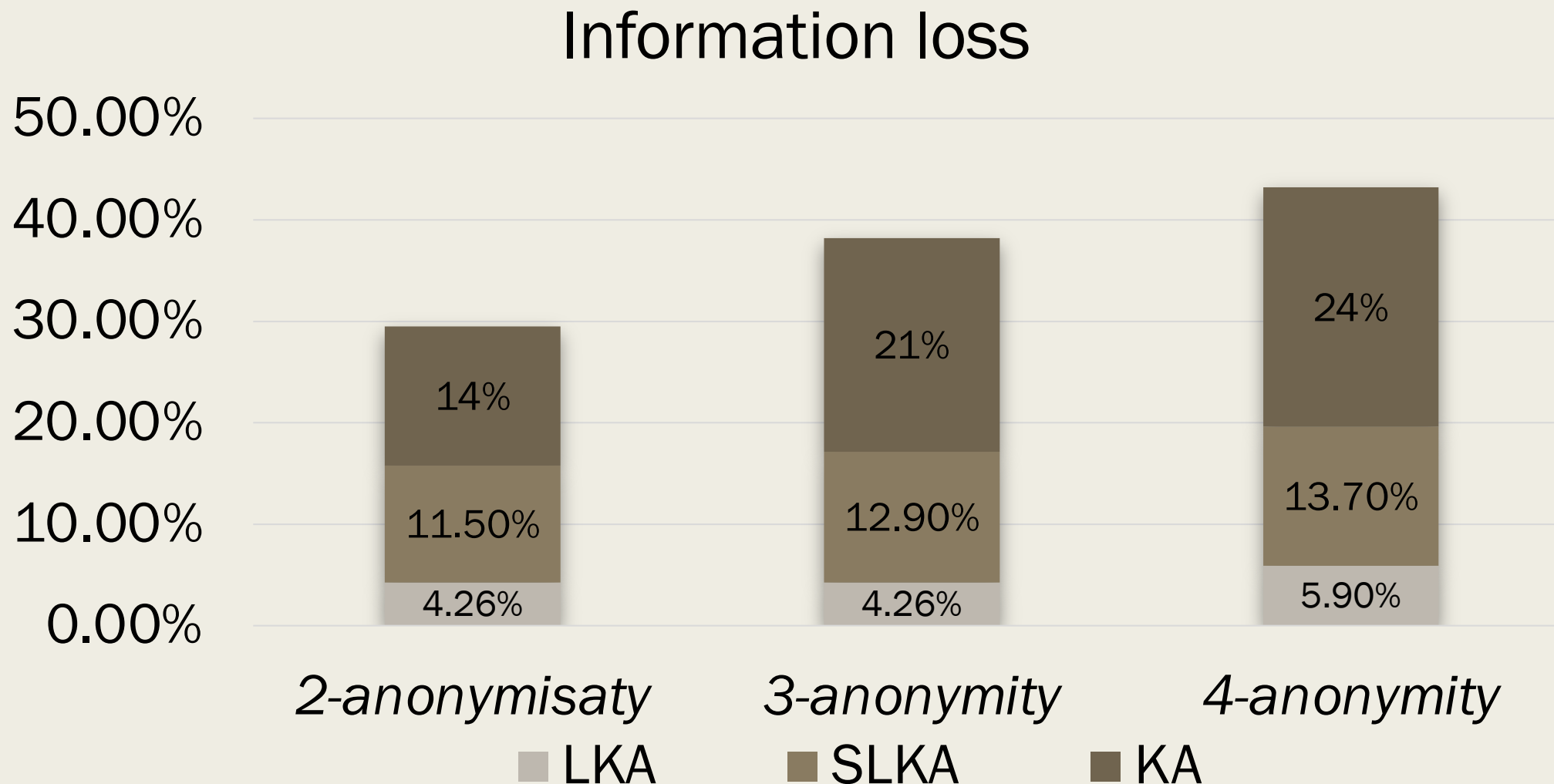


Information Loss (1/2)

$$\text{Information Loss } \frac{\text{SSE}}{\text{SST}} = \frac{\sum_{i=1}^n \sum_{j=1}^m \text{level_Dis}(x_{ij}, x'_{ij})^2}{\sum_{i=1}^n \sum_{j=1}^m \text{level}(x_{ij})^2}$$



Information Loss (2/2)



CONCLUSION

Conclusion

- We adopt semantic web technology to tackle privacy issues...
- *Providing SLKA according to the characteristics of record linkage*
- *Striking the balance between **Privacy** and **Utility***
- *Supporting arbitrary policy (k-anonymity) composition*
- *Improved effectiveness of security policy*

Est.
1841

YORK
ST JOHN
UNIVERSITY

THANK YOU