PRIVACY-PRESERVING ACCESS CONTROL IN ELECTRONIC HEALTH RECORD LINKAGE

Yang Lu
York St John University

Outline

YORK ST JOHN UNIVERSITY Est. 1841

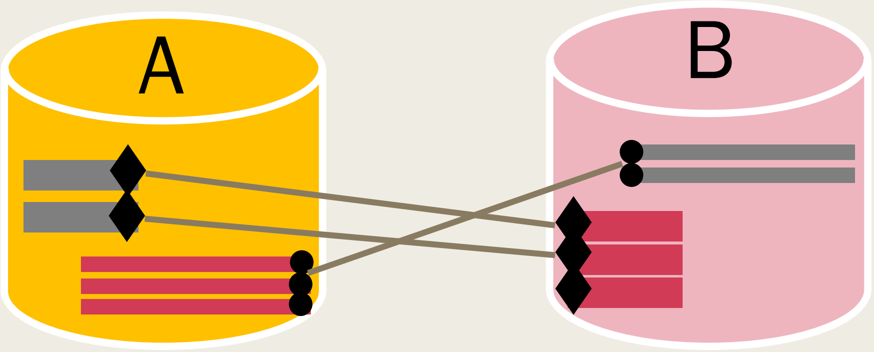
- Background
- Method
- Experiment
- Conclusion

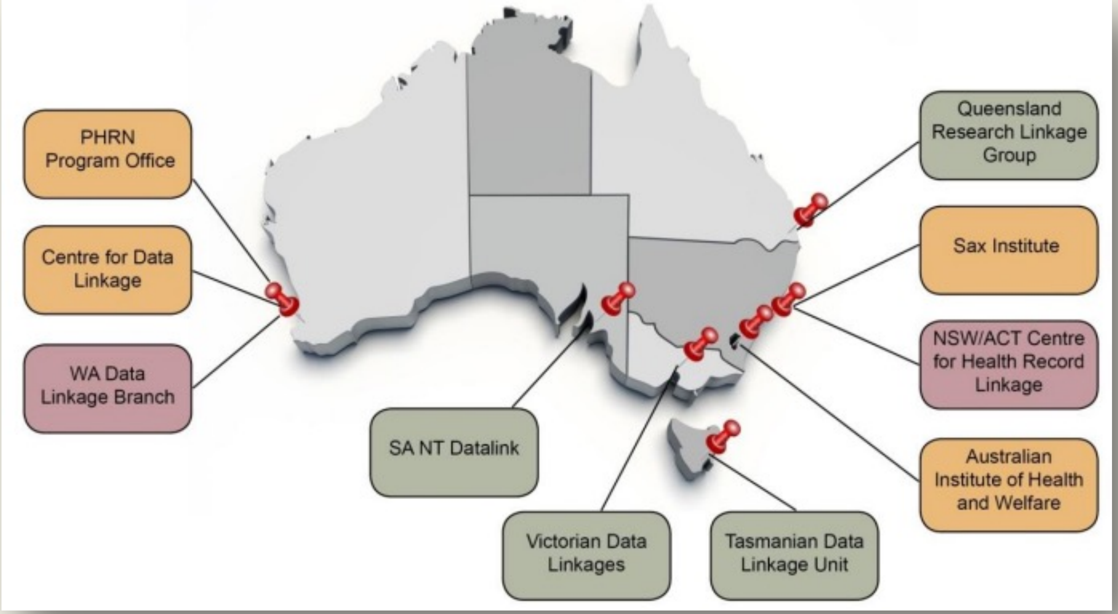
Est. 1841 | YORK ST JOHN UNIVERSITY

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).





Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183-1210.

Protected Personal Information

HIPAA
Health Insurance Portability and Accountability Act

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

Australian Government
National Health and Medical Research Council
NHMRC

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

General Data Protection Regulation

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

Est. 1841 | YORK ST JOHN UNIVERSITY

- “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*

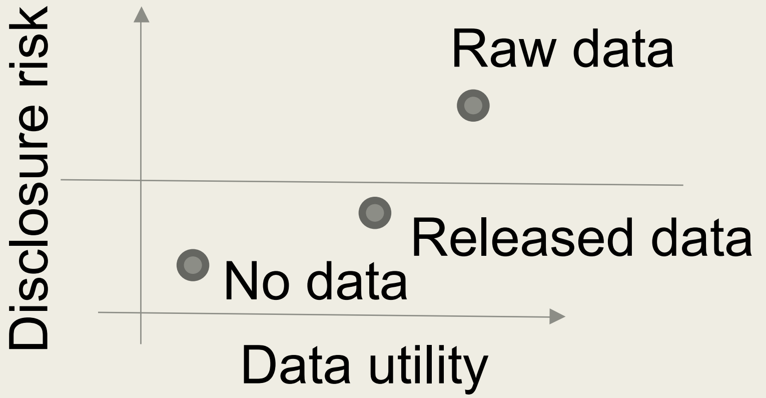
Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco)*, *671*, 1-34.

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	*

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570.

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

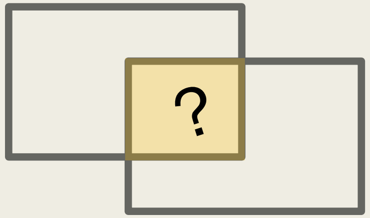
4-anonymity

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	*

The "presence" of individuals is known in k-anonymity case

Linkage disclosure - "presence" is unknow



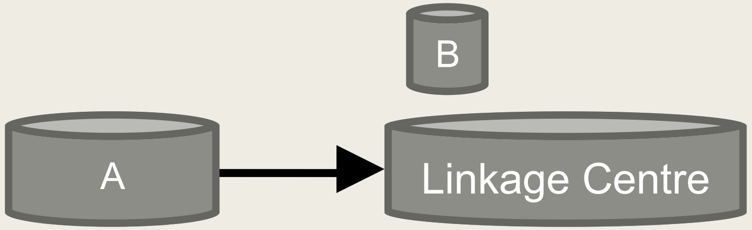
Weak k-anonymity (Atzori, 2006)

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

Atzori, M. (2006, August). Weak k-anonymity: A low-distortion model for protecting privacy. In *International Conference on Information Security* (pp. 60-71). Springer, Berlin, Heidelberg.

Linkage 3-Anonymity (*Time n*)

Est. 1841 YORK ST JOHN UNIVERSITY



(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

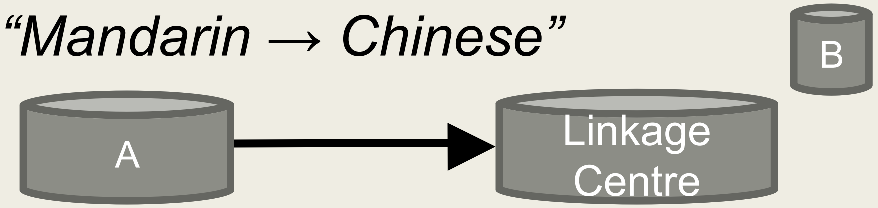
	YoB	Sex	Ethnicity	Language
1	1970-1980	F	Chinese	Mandarin
2	1960-1980	F	Chinese	Mandarin
3	1960-1980	*	Malaysian	English

“Mandarin → Chinese”

Est. 1841 YORK ST JOHN UNIVERSITY

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

"Mandarin → 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

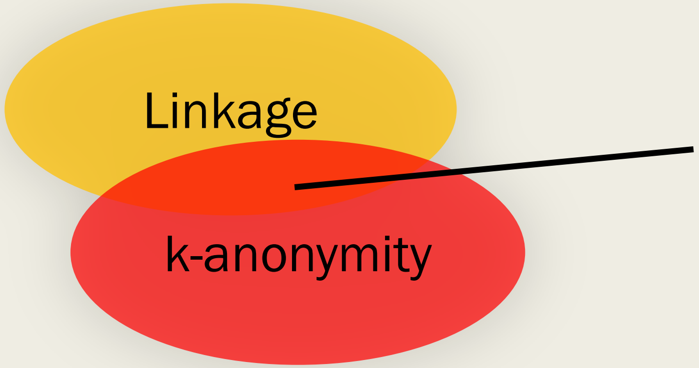
Design Process - SLKA

Est. 1841 YORK ST JOHN UNIVERSITY

Baseline k-anonymity

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

Linkage k-anonymity

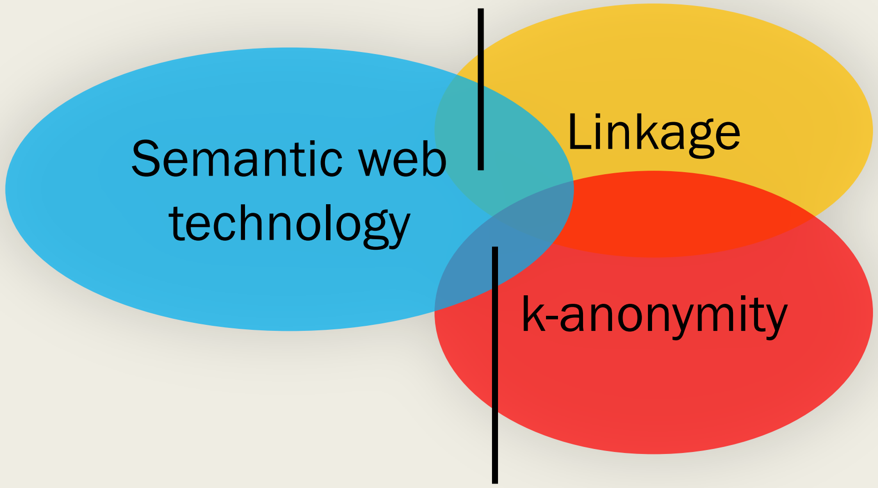


Less information loss

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

Semantic-based linkage k-anonymity (SLKA)

Privacy verification in dynamic datasets



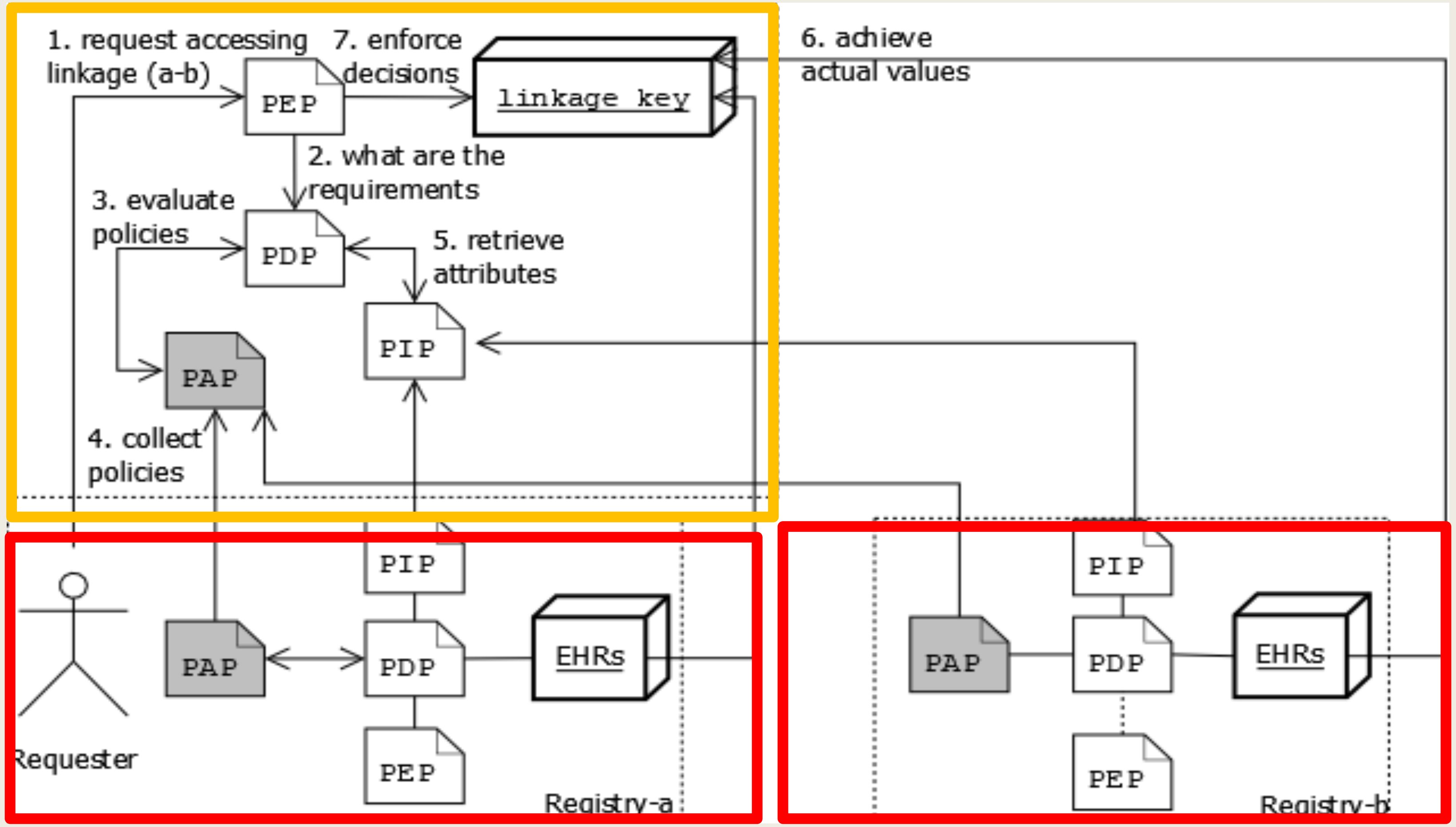
Policy/Rule composition to arbitrary linkage

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

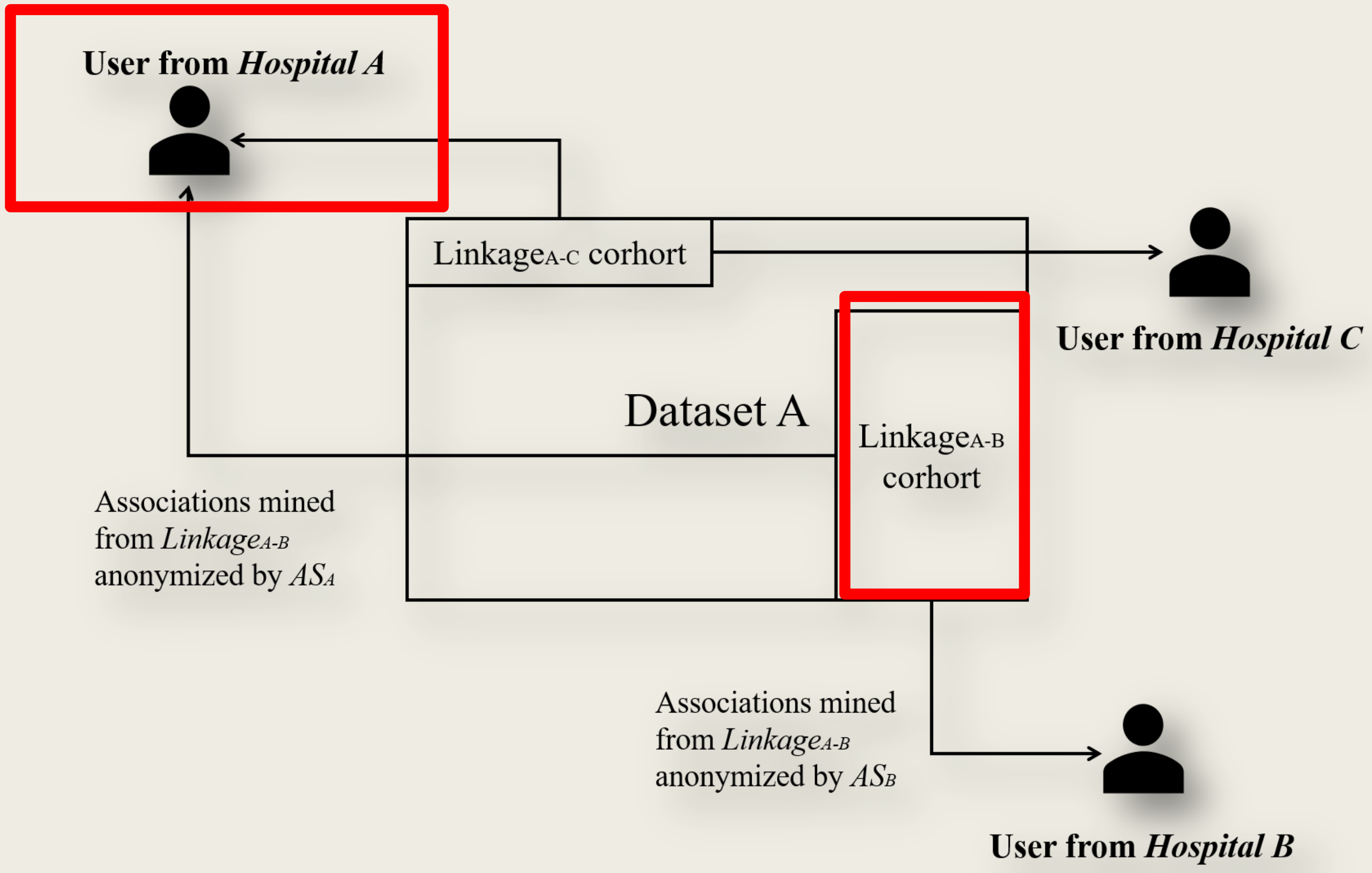
YORK ST JOHN UNIVERSITY Est. 1841

METHOD

Extending XACML Framework

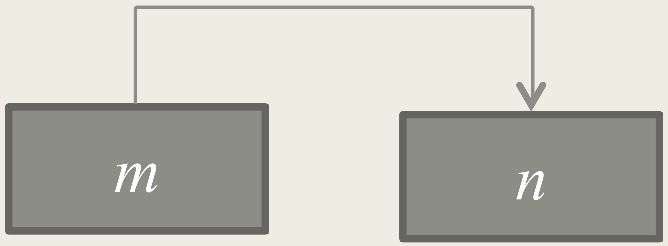


Role-based Knowledge Management

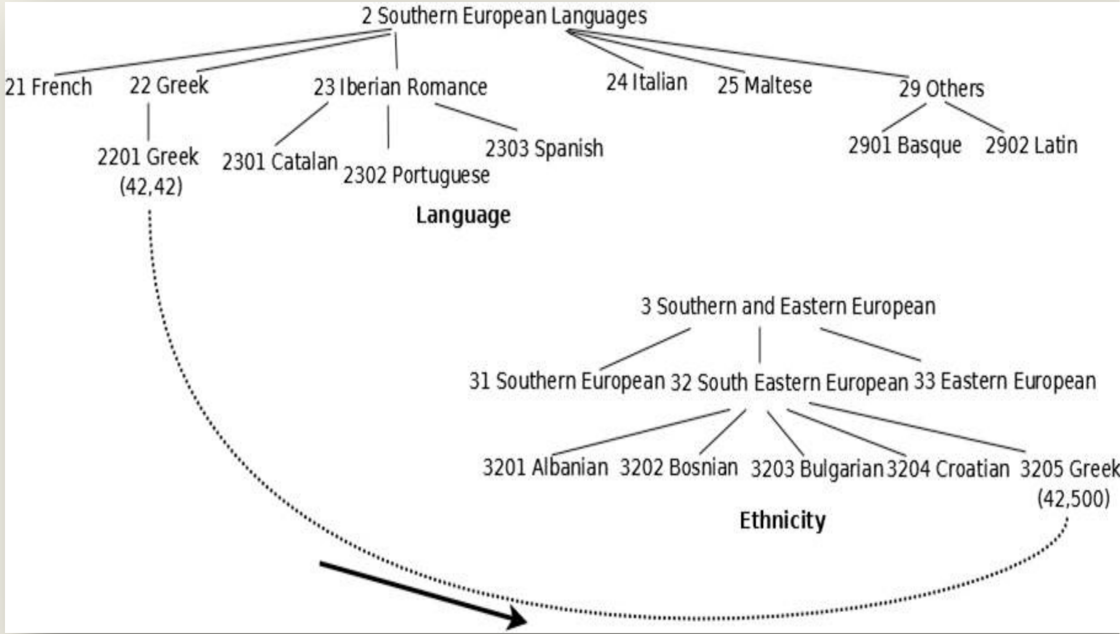


Semantic-based Inference Control

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



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



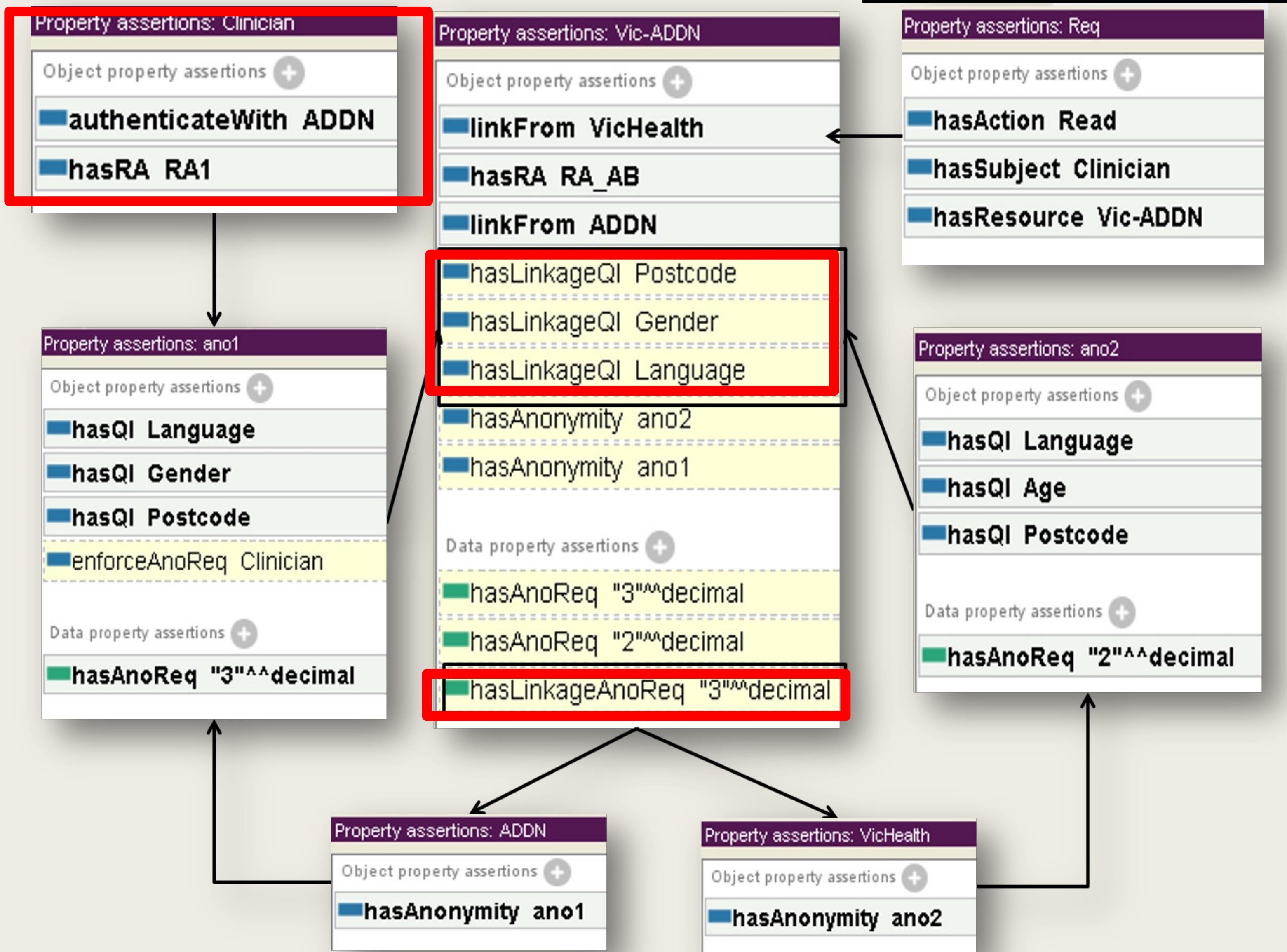
Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

Formalisation and Evaluation (1/2)

Est. 1841 YORK ST JOHN UNIVERSITY

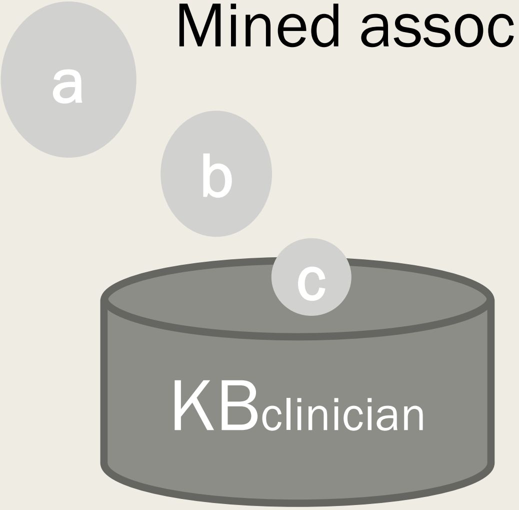
Requestor (role) → Policy scheming

→ Privacy verification



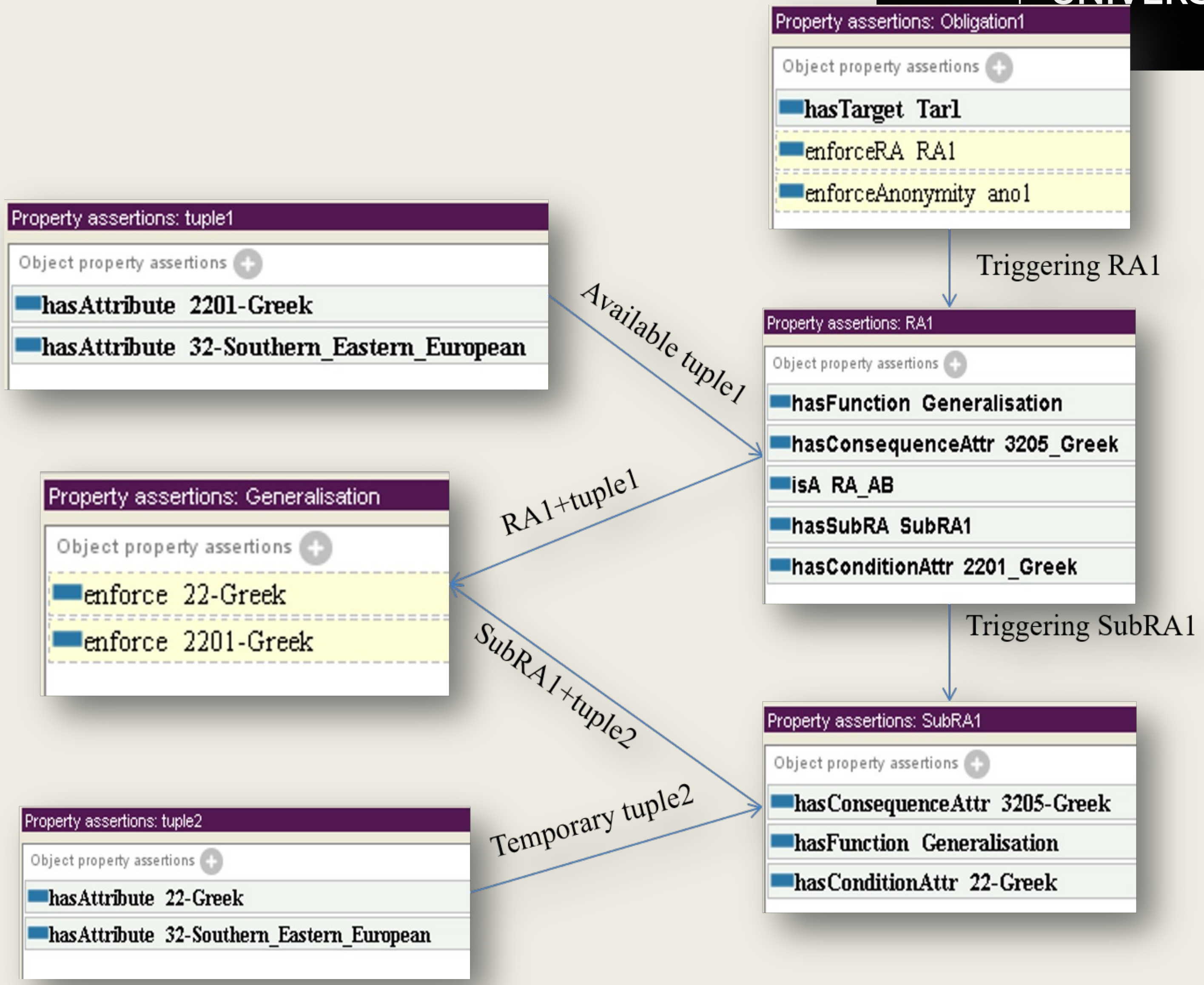
Formalisation and Evaluation (2/2)

Mined associations



NQI -> LQI

e.g. “2201 Greek” -> “3205 Greek”



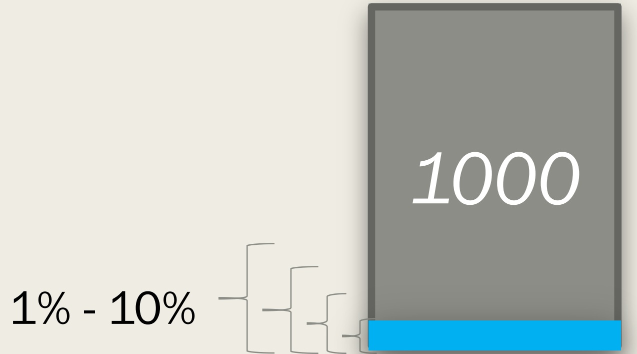
Est. 1841 YORK ST JOHN UNIVERSITY

EXPERIMENT

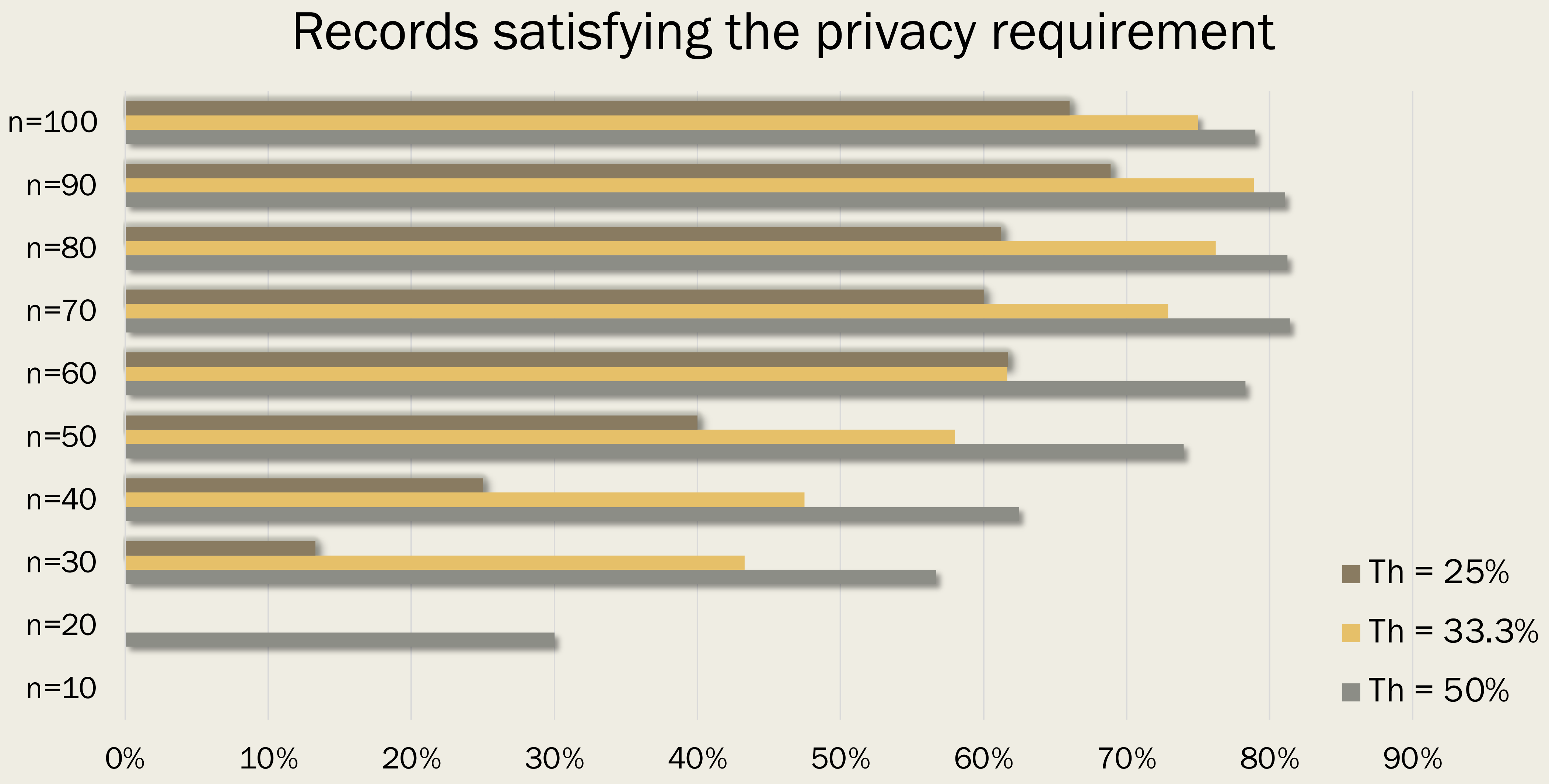
Simulation

Est. 1841 | YORK ST JOHN UNIVERSITY

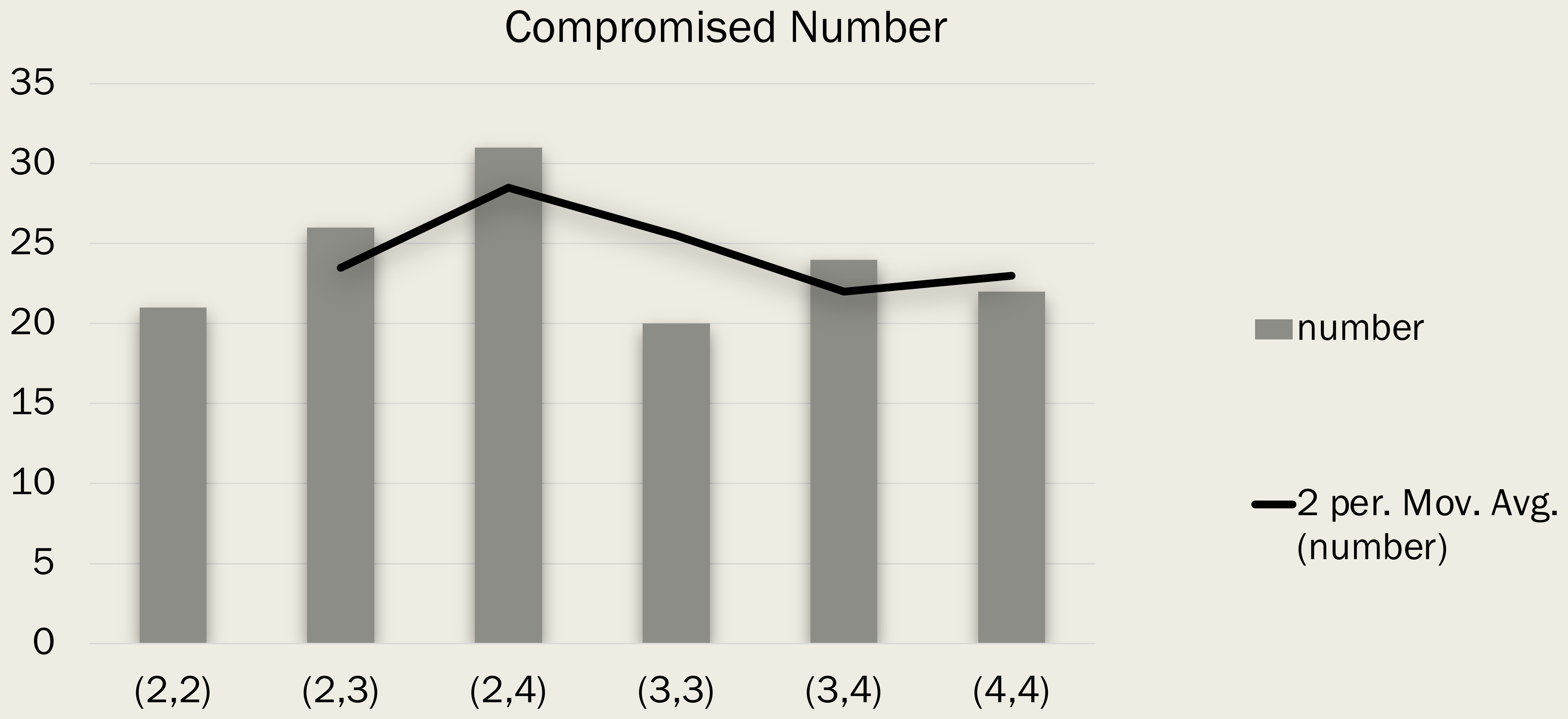
- *1000* (*996* after cleaning) synthetic records about VicHealth survey respondents
- Label records as the "linked" (10 linkage datasets)
- Simulating ***repeated linkage*** requests (Time1 & Time 2)
  - *Same: VicHealth user (role), candidate datasets, cohort, identifiers*
  - *Different: policies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



Security Condition



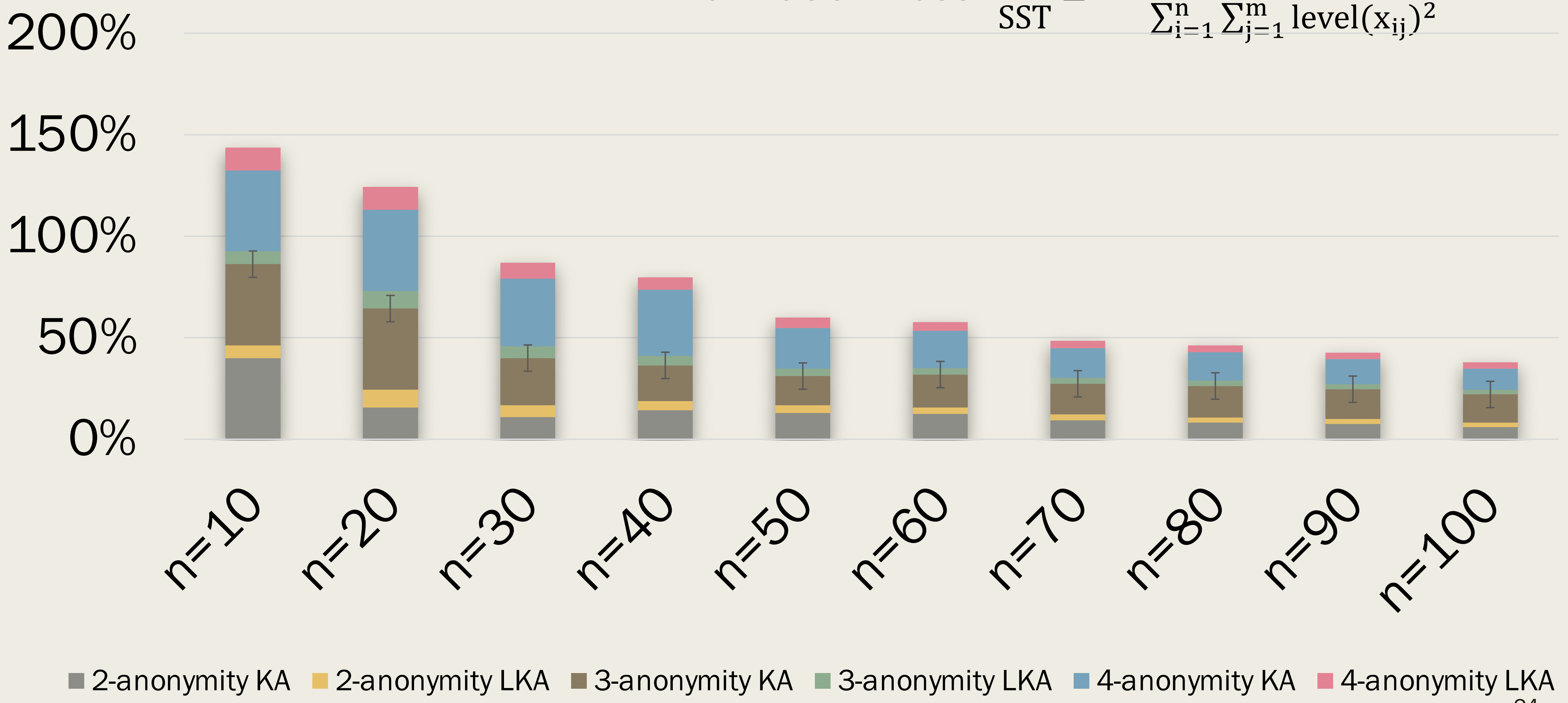
Privacy Violation – Weak k-anonymity



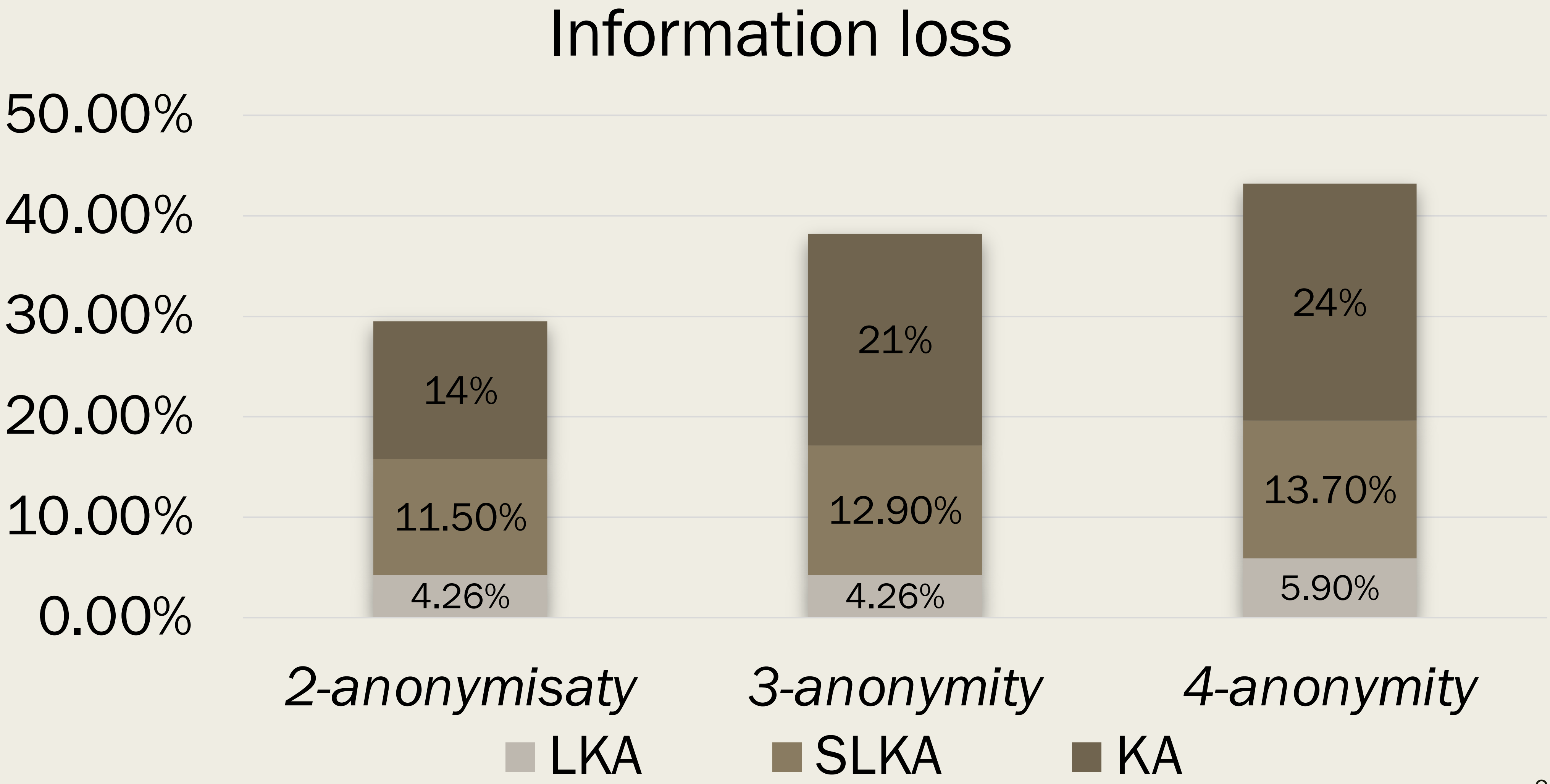
Information Loss (1/2)

Est. 1841 YORK ST JOHN UNIVERSITY

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



Information Loss (2/2)



Est. 1841 | YORK ST JOHN UNIVERSITY

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*

Lu, Y., Sinnott, R. O., Verspoor, K., & Parampalli, U. (2018, August). Privacy-preserving access control in electronic health record linkage. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)* (pp. 1079-1090). IEEE.

Est. 1841 YORK ST JOHN UNIVERSITY

THANK YOU