


On balancing disclosure risk and data utility in transaction data sharing using R-U condentiality map

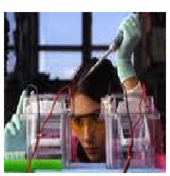
G. Loukides[1] *A. Gkoulalas-Divanis*[2] *J. Shao*[1]

{g.loukides, j.shao}@cs.cf.ac.uk
Cardiff University, UK

agd@zurich.ibm.com
IBM Research – Zurich

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 2011

Individuals’ data is increasingly collected and shared

- **Applications**
  - NETFLIX published movie ratings of 500K subscribers
  - AOL published 20M search query terms of 658K web users
  - TOMTOM sold customers’ location (GPS) data to the Dutch police
  -  eMERGE published patient data related to genome-wide association studies (GWAS) to biorepositories

GWAS associate diseases with DNA – important for personalized medicine

Data sharing is useful

- **<u>Benefits</u>**
  - **Personalization**
    - NETFLIX data mining contest ($1M prize) to improve movie recommendation based on personal preferences
  - **Marketing**
    - TESCO made £53M from selling shopping patterns to retailers and manufacturers (e.g., Nestle and Unilever) last year
  - **Science advancement**
    - Personalized medicine, low-cost social studies

Content

- **Transaction data anonymization**
- **R-U Confidentiality map**
- **Experimental evaluation**
- **Conclusions & future work**

Transaction data

- **A type of data used in many data sharing scenarios**
- **A record (*transaction)* per individual, comprised of a set of items**

Name	*Purchased items*
Mary	a b c d g
Bob	a c e f h i
Tom	b c d g j
Anne	e f g h
Brad	a b d e j
Jim	c f i

Patient	*Diagnosis Codes*
Alice	a, b, c, d, e, f, g, h
Mary	a, c, e, f, g
George	c, d, e, f, h
Jack	a, c, e, f
Anne	e, f, g, h
Tom	d, e, f, g
Jim	a, b, d, e
Steve	a, c, f
David	a, c
Ellen	b, h

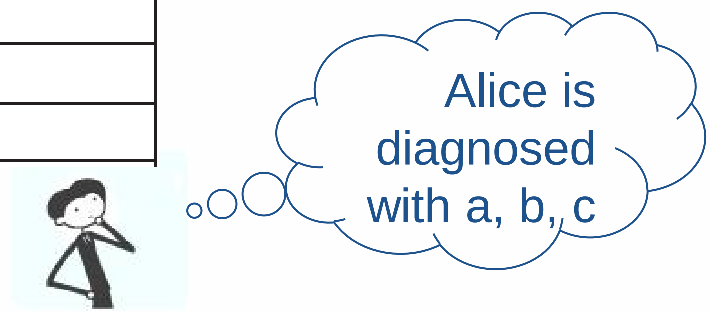
records (transactions)

items

De-identification & identity disclosure

- **<u>Identity disclosure:</u> An individual is linked to her transaction (an attacker learns all her items)**

Patient	*Diagnosis Codes*
Alice	a, b, c, d, e, f, g, h
Mary	a, c, e, f, g
George	c, d, e, f, h
Jack	a, c, e, f



Background knowledge

Alice →

Diagnosis Codes
a, b, c, d, e, f, g, h
a, c, e, f, g
c, d, e, f, h
a, c, e, f

- **Netflix data** – movie rates can be linked to individuals based on IMDB data[1]
- **EMR data** – diagnosis codes can be linked to patients based on public hospital discharge summaries[2]

Data transformation techniques to prevent identity disclosure

- **Item suppression:** Removes items from the published data[2]

Alice has
a, ***b***, and ***c***

Diagnosis Codes
a, b, c, d, e, f, g, h
a, c, e, f, g
c, d, e, f, h
a, c, e, f
e, f, g, h
d, e, f, g
a, b, d, e
a, c, f
a, c
b, h

Suppression

Diagnosis Codes
e, f
e, f
e, f
e, f
e, f
e, f
e
f
-
-

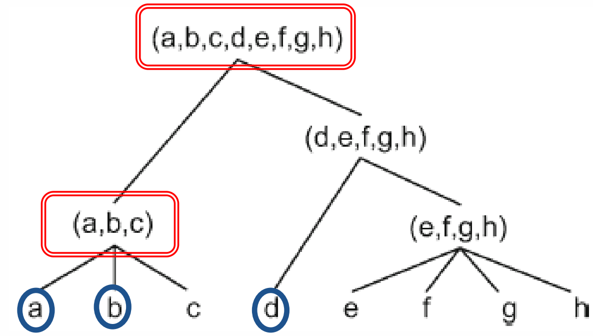
Suppressed result
a, **b**, **c**, **d**, **g**, **h** are not released!

Data transformation techniques to prevent identity disclosure

- **Full-subtree generalization:** Replaces entire *subtrees* of items in a hierarchy with one of their ancestors

Diagnosis Codes
a, b, c, d, e, f, g, h
a, c, e, f, g
c, d, e, f, h
a, c, e, f
e, f, g, h
d, e, f, g
a, b, d, e
a, c, f
a, c
b, h

Alice has
a, ***b***, and ***c***



Full-subtree
Generalization

Diagnosis Codes
(a,b,c), (d,e,f,g,h)
(a,b,c), (d,e,f,g,h)
(a,b,c), (d,e,f,g,h)
(a,b,c), (d,e,f,g,h)
(d,e,f,g,h)
(d,e,f,g,h)
(a,b,c), (d,e,f,g,h)
(a,b,c), (d,e,f,g,h)
(a,b,c)
(a,b,c), (d,e,f,g,h)

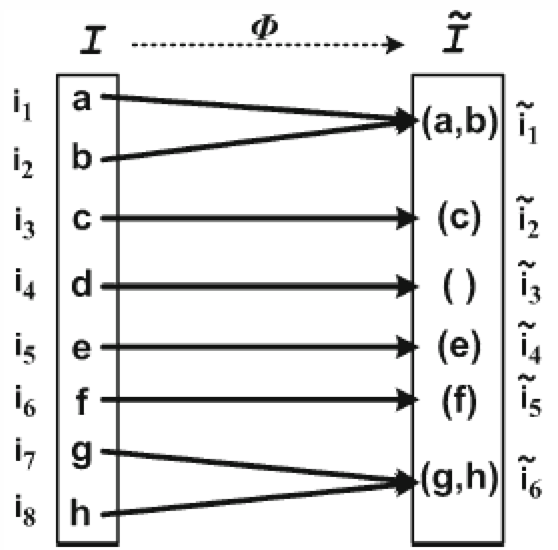
a and ***b*** cannot be generalized **together**

↓

High information loss!

Data transformation techniques to prevent identity disclosure

- **Set-based generalization**: maps items to *generalized* items[3]



Diagnosis Codes
a, b, c, d, e, f, g, h
a, c, e, f, g
c, d, e, f, h
a, c, e, f
e, f, g, h
d, e, f, g
a, b, d, e
a, c, f
a, c
b, h

Alice has
a, ***b***, and ***c***

Set-based
Generalization

Diagnosis Codes
(a, b), c, e, f, (g, h)
(a, b), c, e, f, (g, h)
c, e, f, (g, h)
(a, b), c, e, f
e, f, (g, h)
e, f, (g, h)
(a, b), e
(a, b), c, f
(a, b), c
(a, b), (g, h)

Learn a mapping function ***Φ***
(hierarchies are not necessary)

a and ***b*** are generalized **together**

Balancing data utility with privacy

- **Both suppression and generalization reduce data utility**
  - **Information loss**

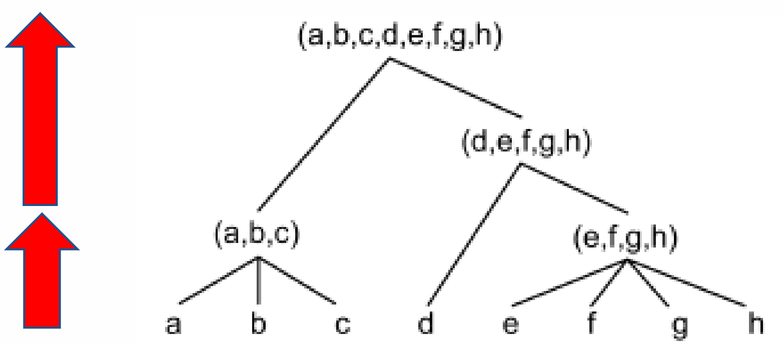
- **Data utility and privacy can only be traded-off**
  - Max utility → Min privacy
  - Max privacy → Min utility

- **Most research so far focused on developing anonymization methods (models and algorithms)**

- **<u>This paper's focus</u>** - *How to use anonymization methods to balance data utility and privacy*

Anonymization principles & algorithms

- **k^m-anonymity:** Knowing that an individual is associated with any m-itemset, an attacker should not be able to associate this individual to less than *k* transactions[4]
- Apriori Anonymization (Rough Sketch)
  - Start with original data
  - While(k^m-anonymity is not satisfied)
    - Generalize items using full-subtree generalization and with minimum information loss
  - Release anonymized data



$Diagnosis\ Codes$
$(a,b,c), (d,e,f,g,h)$
$(a,b,c), (d,e,f,g,h)$
$(a,b,c), (d,e,f,g,h)$
$(a,b,c), (d,e,f,g,h)$
(d,e,f,g,h)
(d,e,f,g,h)
$(a,b,c), (d,e,f,g,h)$
$(a,b,c), (d,e,f,g,h)$
(a,b,c)
$(a,b,c), (d,e,f,g,h)$

- Assumes that the anonymization with the best data utility, for a given privacy requirement (*parameter **m***) needs to be found

Anonymization principles & algorithms

- **Privacy-constrained Anonymity**: Knowing that an individual is associated with one or more privacy constraints *(sets of identifying items)*, an attacker should not be able to associate this individual to less than *k* transactions[5]

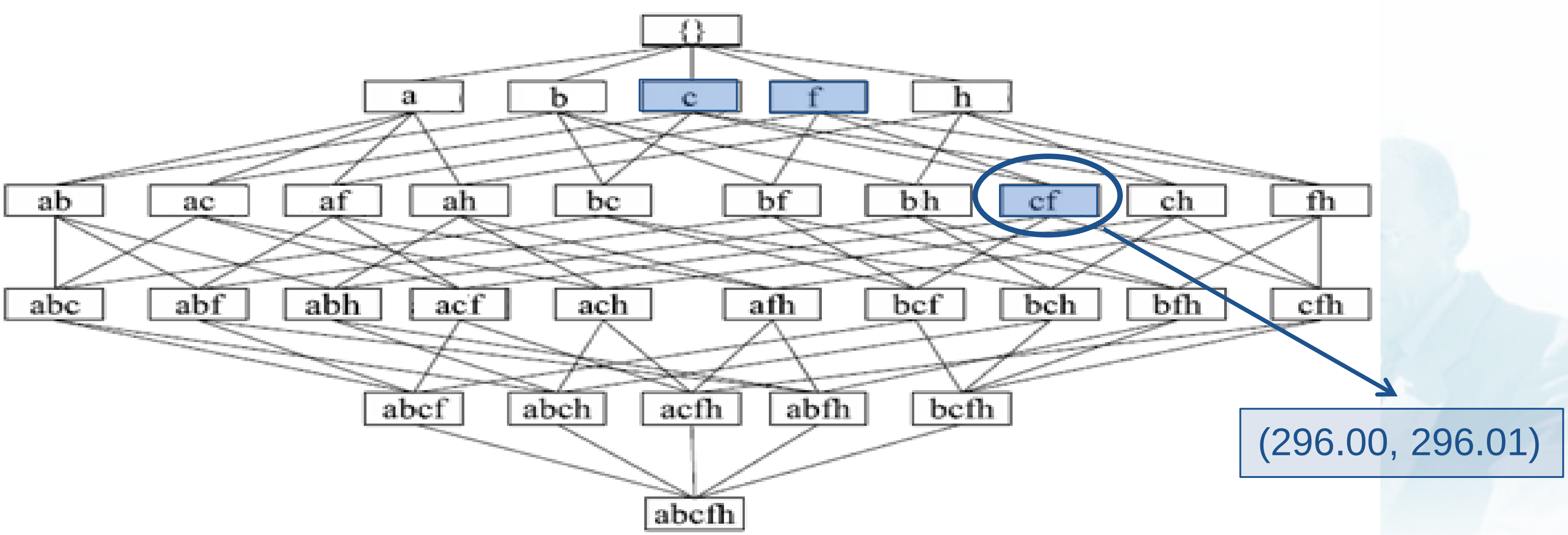
- **PCTA** (Rough Sketch)
  - Start with original data
  - For each privacy constraint
    - While (the privacy constraint is not satisfied)
      - Generalize items using set-based generalization and with minimum information loss
  - Release anonymized data

- Assumes that the anonymization with the best data utility, for a given privacy requirement (*set of privacy constraints*) needs to be found

Anonymization principles & algorithms

- **Privacy and utility constrained anonymity:** Privacy constraints are satisfied; the level of data generalization and suppression is less than what is specified by *utility constraints* (sets of items that are allowed to be mapped to the same generalized item)[3]

- Satisfying utility constraints guarantees data utility in aggregate query answering and in Genome-Wide Association Studies (GWAS)



Anonymization principles & algorithms

- <u>**COAT**</u> (Rough Sketch)
  - Start with original data
  - While (there exists a privacy constraint that is not satisfied)
    - Select the privacy constraint ***p*** that can be protected with minimal information loss
    - While (***p*** is not satisfied)
      - Select the least supported item ***i*** in ***p***
        - **If (*i*** can be anonymized according to the utility constraints)

          generalize i to **(i,i')**
        - **Else**

          suppress items in ***p***, *starting from the least supported item*
  - Release anonymized data

- Assumes that the anonymization with the best data utility, for a given privacy requirement (*set of privacy constraints*) and a given utility requirement (*set of utility constraints*) needs to be found

Utility constraints in Electronic Medical Record data anonymization

Disease	VNEC CBA	VNEC UGACLIP	ACLIP
Asthma	✓	✓	
Attention deficit with hyperactivity	✓		
Bipolar I disorder		✓	
Bladder cancer	✓		
Breast cancer	✓	✓	
Coronary disease		✓	
Dental caries	✓	✓	
Diabetes mellitus type-1		✓	
Diabetes mellitus type-2		✓	
Lung cancer	✓	✓	
Pancreatic cancer	✓	✓	
Platelet phenotypes	✓		
Pre-term birth	✓	✓	
Prostate cancer	✓	✓	
Psoriasis	✓		
Renal cancer	✓		
Schizophrenia	✓		
Sickle-cell disease	✓		

Diseases related to all GWAS conducted until 2008*

no utility constraints

- Result of ACLIP is useless for validating GWAS

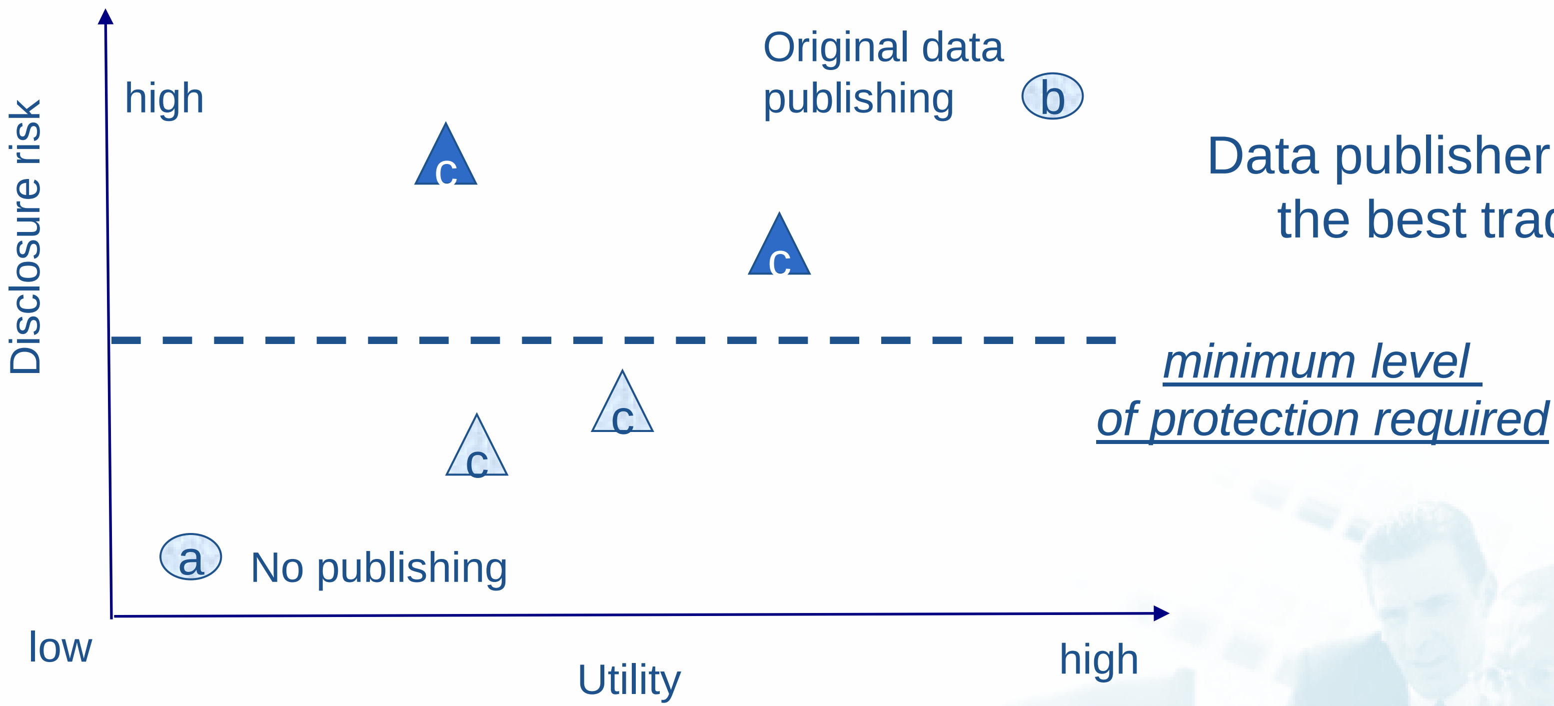
UGACLIP preserves 11 out of 18 GWAS

CBA 14 out of 18 GWAS simultaneously

* Manolio et al. A HapMap harvest of insights into the genetics of common disease. J Clinic. Inv. '08.

Tracking the utility/privacy trade-off

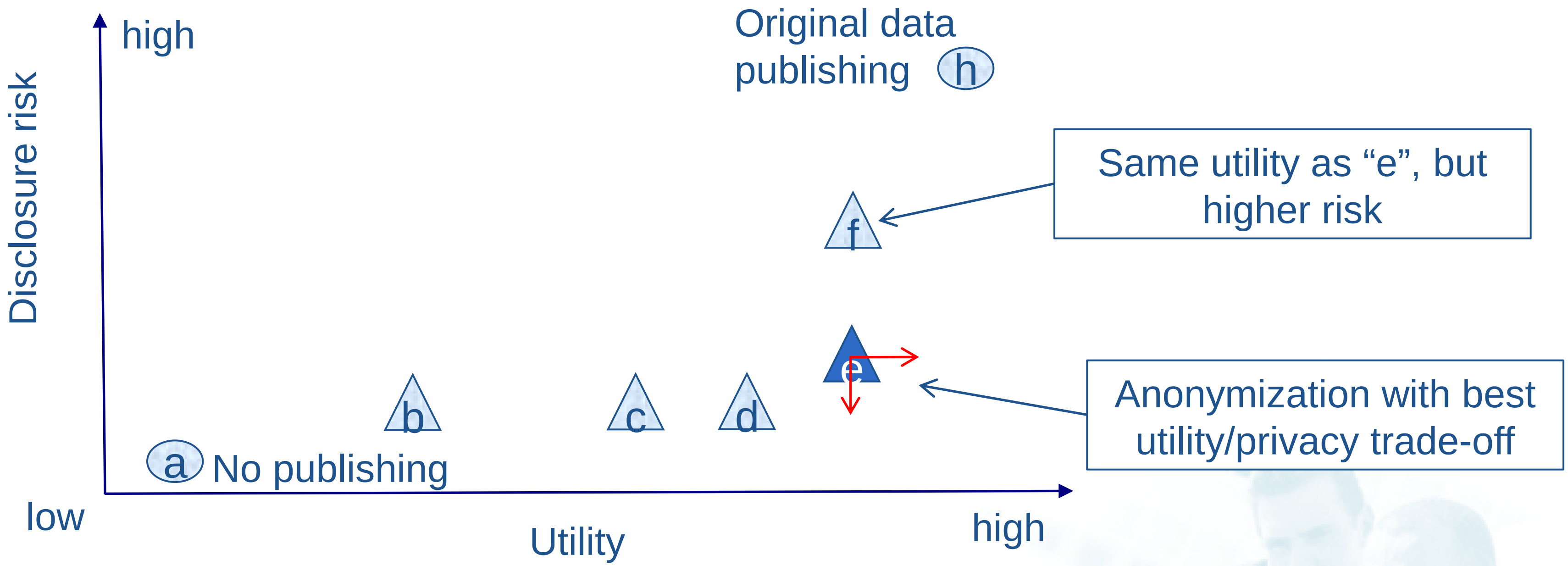
- **R-U confidentiality map[6]**



- Proposed for additive noise, applied to k-anonymization and randomization[7]
- **What does it offer?**
- **Can it be used for transaction data anonymization?**

Tracking the utility/privacy trade-off

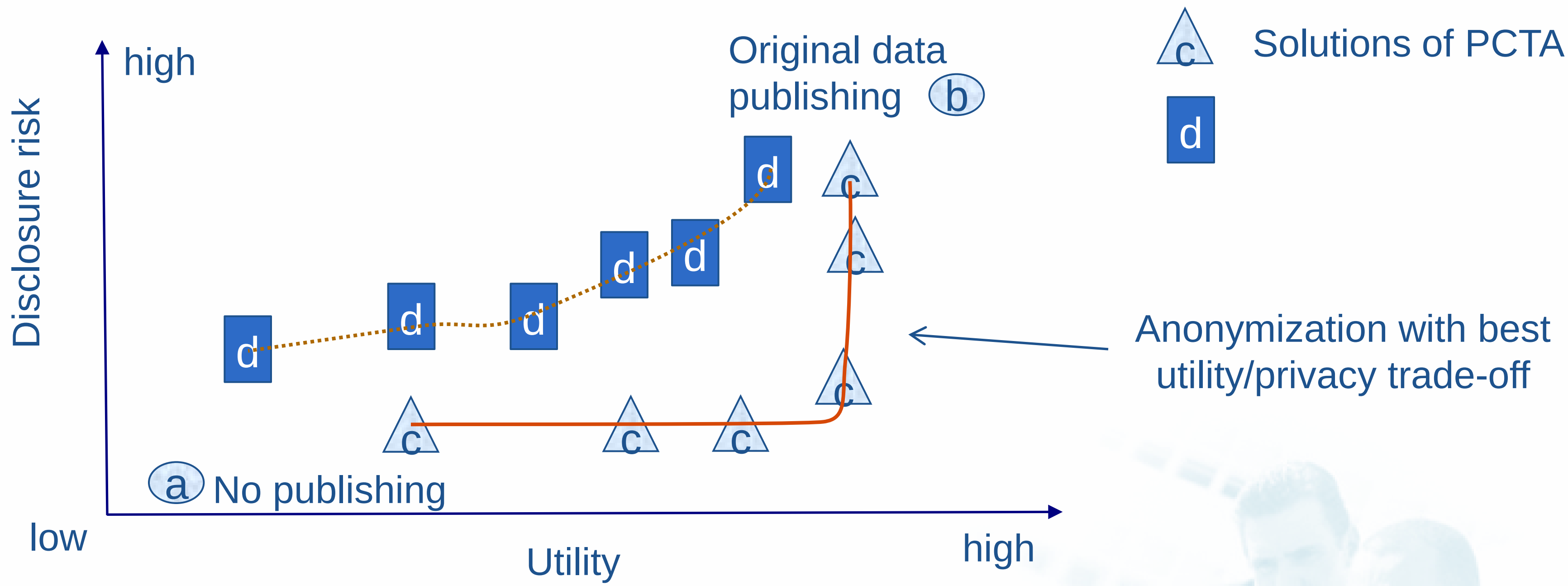
- **R-U confidentiality map[6]**



- **Data publishers attempt to configure an anonymization method**
  → R-U confidentiality map can help them find a solution with the best utility/privacy trade-off

Tracking the utility/privacy trade-off

- **R-U confidentiality map[6]**



- **Data publishers want to select an anonymization method to use**
  →R-U confidentiality map allows comparing different methods

Applying R-U confidentiality map to transaction data anonymization

- **A measure for disclosure risk**
  - *Risk* - inverse of the maximum probability identity disclosure occurs

$$\frac{1}{\min_{\forall p \in P} \sup(p, \tilde{D})}$$

privacy constraints

anonymized dataset

- **A measure for Utility**
  - *Utility* - inverse of the Average Relative Error (ARE) $\frac{1}{ARE}$
  - ARE – average number of transactions retrieved *incorrectly*, when answering a workload of queries on anonymized data

$$\texttt{SELECT COUNT}(T_n)\ \texttt{FROM}\ D \quad \texttt{WHERE}\ I \text{ supports } T_n \text{ in } D$$

Query on original data

$$\texttt{SELECT COUNT}(\tilde{T}_n)\ \texttt{FROM}\ \tilde{D} \quad \texttt{WHERE}\ \tilde{I} \text{ supports } \tilde{T}_n \text{ in } \tilde{D}$$

Query on anonymized data

Experimental evaluation

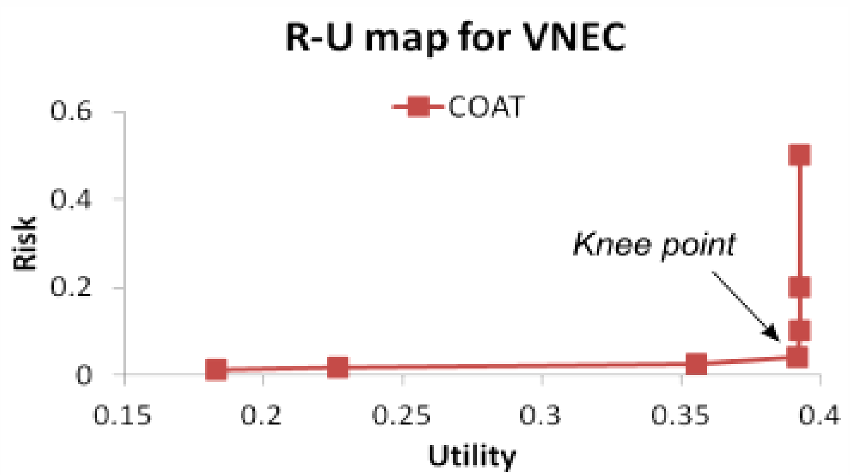
- **Datasets**
  - **BMS-WebView 2 (BMS2)** - click-stream data from an e-commerce site[7]
  - **VNEC** – Electronic Medical Record dataset from Vanderbilt – contains the diagnosis codes of patients involved in a GWAS[8]
  - **VNEC**$_{kc}$ - subset of VNEC, we know which diseases are controls for others[9]

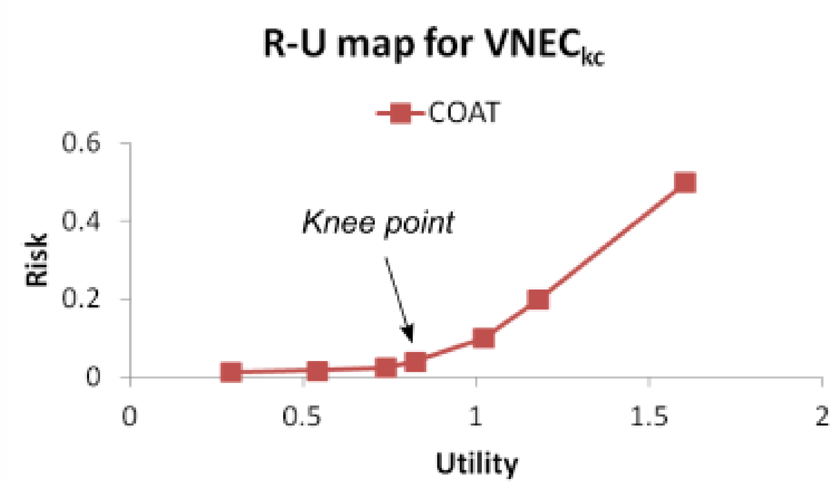
- **Algorithms - **Apriori, COAT, PCTA

- **We constructed R-U maps for**
  - Privacy and utility-constrained anonymity
  - k^m-anonymity

Identify anonymizations with best utility/privacy trade-off

- **Medical datasets VNEC and VNEC$_{kc}$** - COAT algorithm
  - privacy constraints to prevent attacks using hospital discharge summaries [8]
  - utility constraints to guarantee utility for 18 Genome-Wide Association Studies [8]

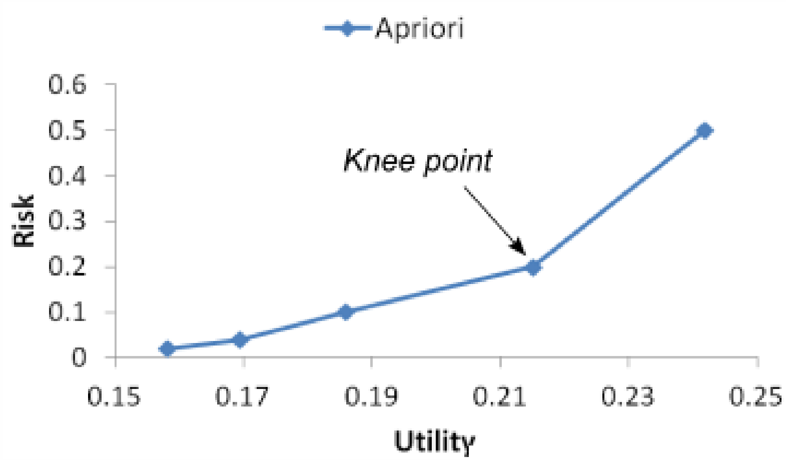


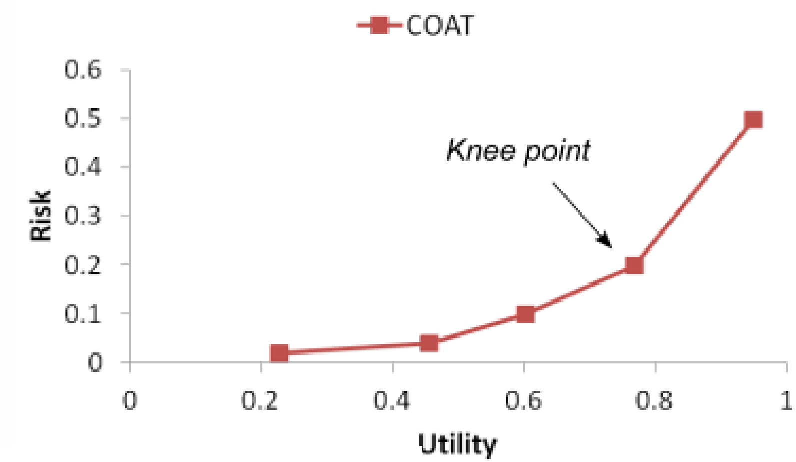


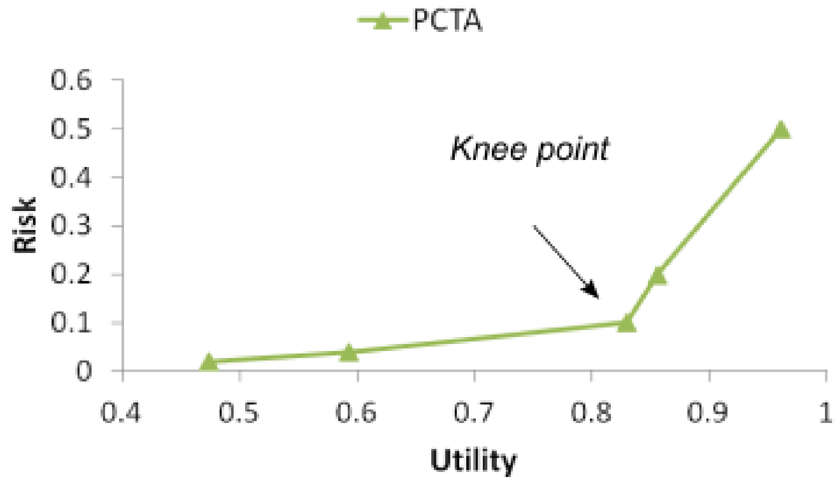
  - Knee point corresponds to the anonymization with “best” trade-off, found by the Angle-based method [9]

Find anonymizations with best utility/privacy trade-off

- **BMS2 dataset –** k^2-anonymity & Apriori, COAT, PCTA algorithms



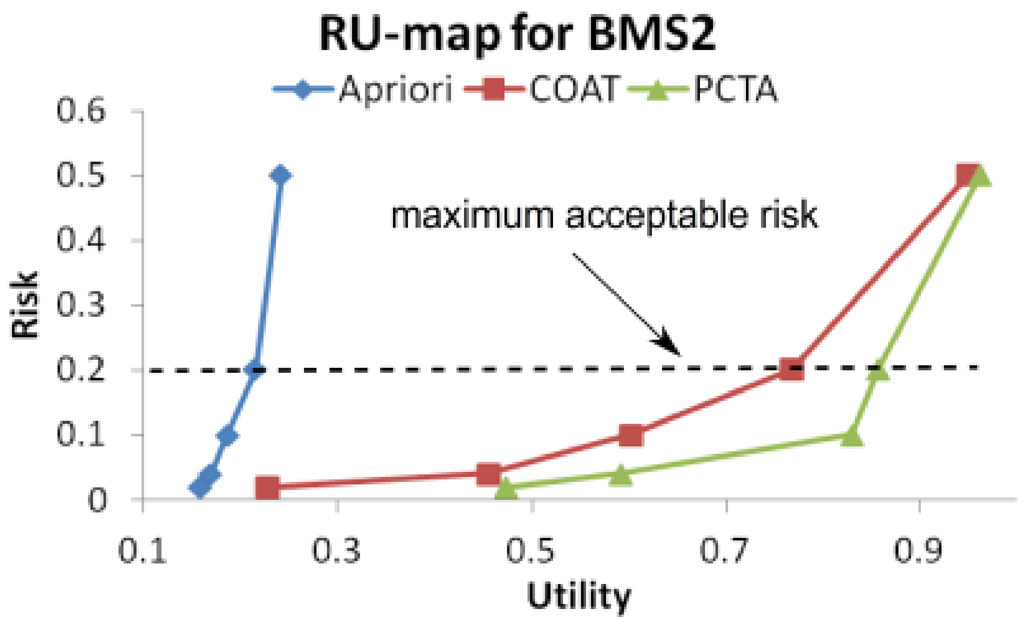




  - Knee point corresponds to the anonymization with "best" trade-off, found by the Angle-based method [9]

Select anonymization method, given a maximum level of *Risk*

- **BMS2 dataset –** k^2-anonymity & Apriori, COAT, PCTA algorithms
  - Data publishers want to release anonymized data with Risk no more than 0.2



  - They should use PCTA, because it produces anonymized data with higher *Utility* when *Risk* is 0.2 or less.

Conclusions & future work

- **Need for publishing transaction data**
- **Several recent methods for anonymizing transaction data**
- **How to trade-off data utility and privacy using R-U map**
- **In the future**
  - **Apply R-U map to compare methods using different privacy models**
    - Generalization vs. noise addition
  - **Different ways to balance data utility and privacy**
    - Methods that optimize the utility/privacy trade-off

References & Acknowledgements

1. A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy, 2008.
2. Y. Xu et al. Anonymizing transaction databases for publication. ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2008.
3. G. Loukides, A. Gkoulalas-Divanis and B. Malin. COAT: Constraint-based Anonymization of Transactions. Knowledge and Information Systems: An International Journal (KAIS).
4. M. Terrovitis, N. Mamoulis, and P. Kalnis. Local and Global Recoding Methods for Anonymizing Set-valued Data. VLDB Journal, 2010.
5. A. Gkoulalas-Divanis and G. Loukides. Privacy-constrained Clustering-based Transaction Data Anonymization. EDBT International Workshop on Privacy and Anonymity in the Information Society.
6. G.T.Duncan, S.A.Keller-McNulty, and S.L. Stokes. Disclosure risk vs. data utility: The R-U Confidentiality Map. Los Alamos National Library Technical Report, LAUR-01-6428. 2001.
7. Z. Zheng, R. Kohavi, and L. Mason. Real world performance of association rule algorithms. In KDD, pages 401{406, 2001.
8. G. Loukides, A. Gkoulalas-Divanis, and B. Malin. Anonymization of Electronic Medical Records for Validating Genome- Wide Association Studies. Proceedings of the National Academy of Sciences, 2010.
9. Q. Zhao, V. Hautamaki, and P. Frnti. Knee point detection in BIC for detecting the number of clusters. In Advanced Concepts for Intelligent Vision Systems, pages 664-673, 2008.

- Dr. Loukides' research is partly supported by

The Royal Academy of Engineering