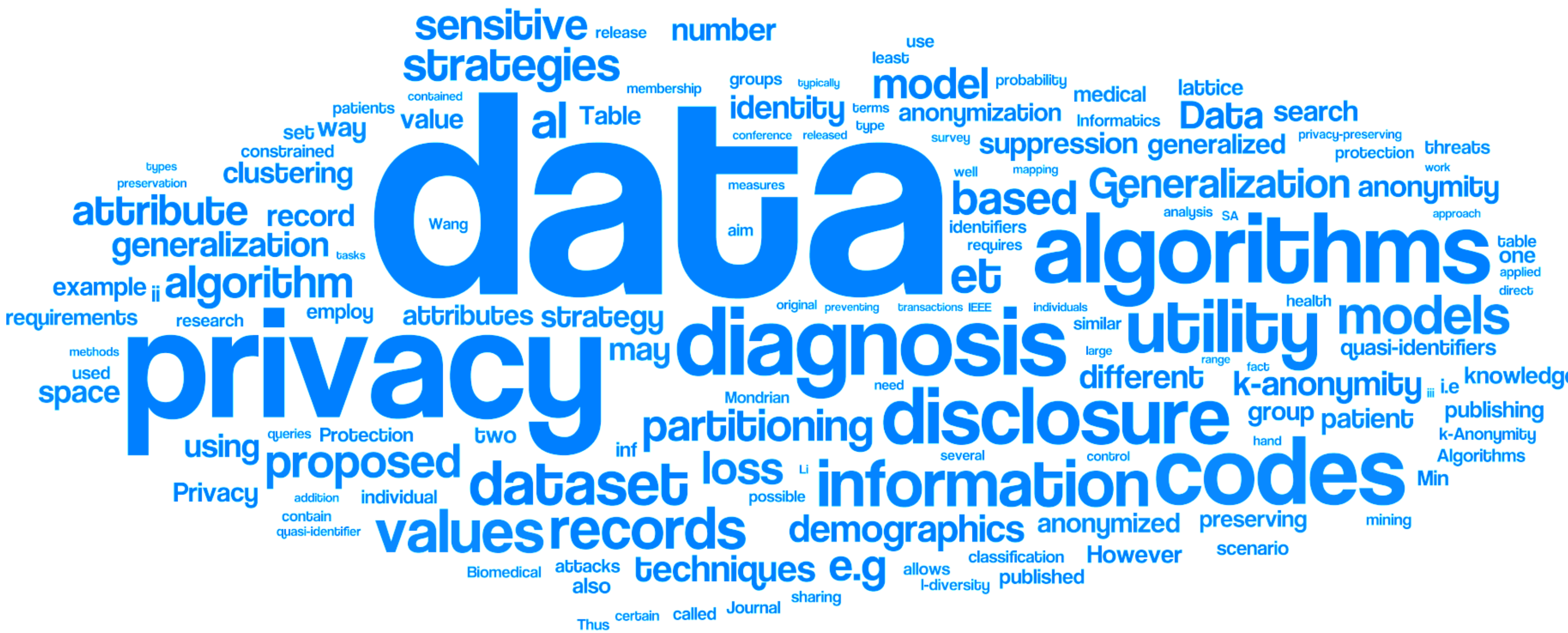
TMF Workshop: Anonymization tools and their practical relevance (for biomedical research)

An overview of state-of-the-art methods



Fabian Prasser

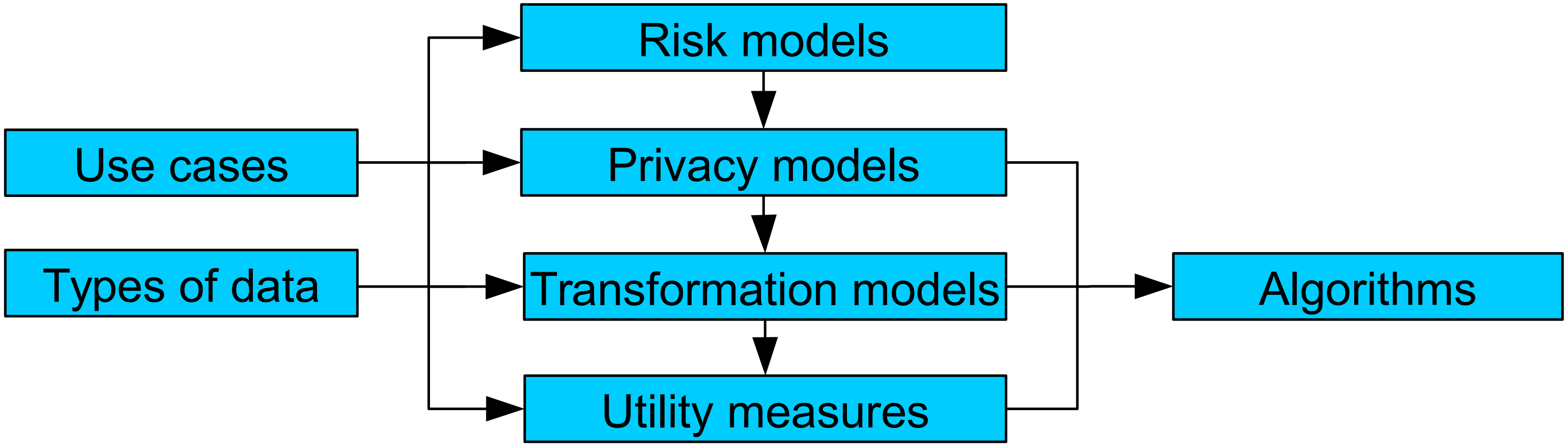
Lehrstuhl für Medizinische Informatik
Institut für Medizinische Statistik und Epidemiologie
Klinikum rechts der Isar der TU München

Three types of de-identification / anonymization

- **Masking identifiers in unstructured data**
  - **Subject:** clinical notes, …
  - **Methods:** machine learning, regular expressions, …
  - **Implementations:** MIST, MITdeid, NLM Scrubber
- **Privacy preserving data analysis (interactive scenario)**
  - **Subject:** query results, …
  - **Methods:** interactive differential privacy, query-set-size control, …
  - **Implementations:** AirCloak, Airavat, Fuzz, PINQ, HIDE
- **Transforming structured data (non-interactive scenario)**
  - **Subject:** tabular data, …
  - **Methods:** generalization, suppression, randomization, ...
  - **Implementations:** AnonTool, ARX, sdcMicro, µArgus, PARAT

Multiple aspects have to be balanced

- **Main goal:** Achieve a balance between data utility and privacy
- **Complex task**
  - Many different types of methods need to be applied in an integrated manner
  - Methods may need to be parameterized
  - Different aspects are interrelated
- **Just the most important aspects and relationships**



Important aspects of use cases

- **Who or what will process the data in which way?** [FWF11]
  - **Humans, e.g., epidemiologists**
    - Different types of analyses
    - Interactive vs. non-interactive
  - **Machines, i.e., data mining**
    - Classification vs. clustering
- **How will the data be released?** [FWF11]
  - **Access control**
    - Open access vs. restricted access
  - **Continous data publishing**
    - Multiple views vs. re-release (incremental vs. new attributes)
- **Is the data distributed?** [FWF11]
  - **Collaborative environments**
    - Vertical vs. horizontal vs. hybrid distribution

Important properties of data

- **Relational data**
  - Tabular data
  - One row per individual
- **Transactional data**
  - Data consisting of set-valued attributes
  - Example: Follow-up collection of diagnosis codes
- **Data with relational and transactional characteristics**
- **Dimensionality of data**
  - Mitigating re-identification is practically infeasible for high-dimensional data [Agg05]
- **Data with clusters**
  - Example: household structures
- **Other types of data:** Trajectory data, social network data

Privacy models: some background

- **Definition of (perfect) privacy** [Dal77] formulated by [Dwork08]
  - "*Anything that can be learned about a respondent from a statistical database should be learnable without access to the database*"
- **Syntactic models**
  - Syntactic conditions on the released datasets
  - No (direct) semantic implications regarding the above definition
  - **Instead:** Assumptions about attack vectors and definition of (likely) background knowledge and goals by classifying attributes [Swe02]
    - Direct and indirect identifiers (or quasi-identifiers, or keys)
    - Sensitive and insensitive attributes
- **Semantic models**
  - Privacy models that relax a formalization of the above definition
  - Much fewer assumptions need to be made about attackers

Risk and threat models

- **Disclosure models** [LLZ+12]
  - Identity disclosure (re-identification, tuple linkage)
  - Attribute disclosure (sensitive information disclosure)
  - Membership disclosure (table linkage)
- **Models for quantifying re-identification risks**
  - **Super-population models:** Population is modeled with probability distributions parameterized with sample characteristics
  - **Decision rule by Dankar et al.:** Combination of three models, which has been evaluated for biomedical datasets [DEN+12]
- **Attacker models:** May be used to derive/compile global risks [Emam13]
  - **Prosecutor scenario:** Targets one specific individual
  - **Marketer scenario:** Targets as many individuals as possible
  - **Journalist scenario:** Targets any individual

Syntactic models against re-identification

- **Goal:** Prevent linkage attacks on quasi-identifiers
- **Some models for relational data**
  - **k-Anonymity:** Requires groups (cells or equivalence classes) of size ≥ k, which defines an upper bound on the re-identification risk (over-) estimated with sample frequencies [Swe02]
  - **LKC-Privacy:** Relaxed variant of k-anonymity + (ℓ-diversity) [MFH+09]
  - **Risk-based approaches:** Enforce thresholds on re-identification risks, which may be quantified with super-population models
  - **HIPAA Safe Harbor:** Heuristic with many predefined identifiers and a few quasi-identifiers (regions and all kinds of dates). Contains wildcards („*any other unique identifying number, characteristic, or code*“). Provides sound legal protection for custodians in the US [HIP]
- **Some models for transactional data**
  - **(k^m)-Anonymity:** k-Anonymity regarding ≤ m values from a set [TMK08]

Syntactic models against attribute disclosure

- **Observation:** Preventing linkage attacks is not enough
- **Goal:** Prevent knowledge gain from sensitive information associated with an equivalence class
- **Some models for relational data**
  - **ℓ-Diversity:** Sensitive values must be „well-represented". Multiple variants exist with different privacy/utility trade-offs [MGK+06]
  - **t-Closeness:** Distribution of sensitive values must not be „too different" from the overall dataset. Multiple variants exist [LLV07]
  - **p-Sensitive k-anonymity:** Focus on identity & attribute disclosure [TV06]
  - **LKC Privacy:** ℓ-Diversity & relaxed k-anonymity [MFH+09]
- **Some models for transactional data**
  - **(h, k, p)-coherence:** k^m-anonymity + protection against inference [XWF+08]
  - **ρ-Uncertainty:** protection against inference with fewer assumptions [CKR+10]

Further syntactic models

- **Models against membership disclosure for relational data**
  - **Goal:** Bounds on the certainty with which the presence of data about an individual in a database can be inferred via linkage
  - **Upside:** With strict thresholds, they provide semantic privacy
  - **Downside:** Basically impossible to achieve
  - **δ-Presence:** Relates sample counts to population counts [NAC07]
  - **c-Confident δ-presence:** Relaxation of δ-presence in which population characteristics are estimated [NC10]
- **Models for data which is relational and transactional**
  - **(k, k^m)-Anonymity:** Mixture of k-anonymity and k^m-anonymity [TMK08]
- **Models for continuous publishing of relational data**
  - **Approach by Byun et al.:** Only supports insertions
  - **m-Invariance:** Supports insertions, deletions, updates [XT07]

A semantic model: Differential Privacy

- **Observation** [Dwork06]
  - The formal notion of privacy is impossible to achieve
  - Even for individuals that are not part of the statistical database
- **Idea** [Dwork 06, Dwork08]
  - Do not compare an attacker's information about an individual before and after accessing a statistical database, but
  - Compare the risks for an individual when joining (or leaving) a statistical database
- **(Slightly) more formal** [Dwork 06, Dwork08]
  - **ϵ-Differential Privacy:** A (randomized) function fulfills ϵ-DP if the probability of every possible output value changes by a factor of at most exp(ϵ) when data about an individual is or is not contained in a database.
  - **Relaxations:** (ϵ, δ)-DP, approximate DP [PK08, LM12]

A semantic model: Differential Privacy (cont'd)

- **DP in interactive scenarios** [FDE13]
  - Sequential composition rule
  - Privacy budget
- **DP in non-interactive scenarios**
  - Release of contingency tables or marginals [BCD+07]
  - Relationships to syntactical models exist, e.g., [LQS11]
    - **(k, β)-SDGS:** Random sampling + k-anonymity fulfills (ϵ, δ)-DP
    - **t-Closeness (with a specific distance function):** Implies ϵ-DP regarding the sensitive attributes [DS15]
- **Has been criticized in the context of biomedical research** [FDE13]
  - **DP is often not a truthful mechanism:** Functions are randomized, often data is pertubated, e.g., by adding noise
  - **DP is not intuitive:** What is a good value for ϵ? What does it mean?

Measuring data utility

- **Often used interchangeably with "loss of information"**
- **Exemplary utility measures for syntactic models**
  - Used for evaluating transformed datasets
  - **Discernibility:** Based on sizes of equivalence classes [BA05]
  - **Average equivalence class size:** Analogously to discernibility [LDD+05]
  - **(Non-uniform) entropy:** Information theoretic measure [GT09]
  - **Loss:** Measures the coverage of the domain of attributes [VI02]
  - **Utility constraints:** Use cases are modeled as queries [LGM10]
- **Exemplary utility measures for Differential Privacy** [FDE13]
  - Used for evaluating a method that fulfills DP
  - **Error:** Absolute, relative, variance
  - **(α, δ)-Usefulness:** $P[\text{distance} \leq \alpha] \geq \delta$

Transformation methods

- **Coding models** [FWF11]
  - **Global recoding:** Similar transformation for similar values
  - **Local recoding:** Different transformations may be applied
- **Truthful transformations** [FWF11]
  - **Generalization:** Based on domain generalization hierarchies
    - Full-domain generalization: All values of an attribute are generalized to the same level
    - Subtree generalization: Different levels of generalization may be applied
  - **Suppression:** Removal of values of cells or complete tuples
  - **Top & bottom coding:** Replacing values that exceed given bounds

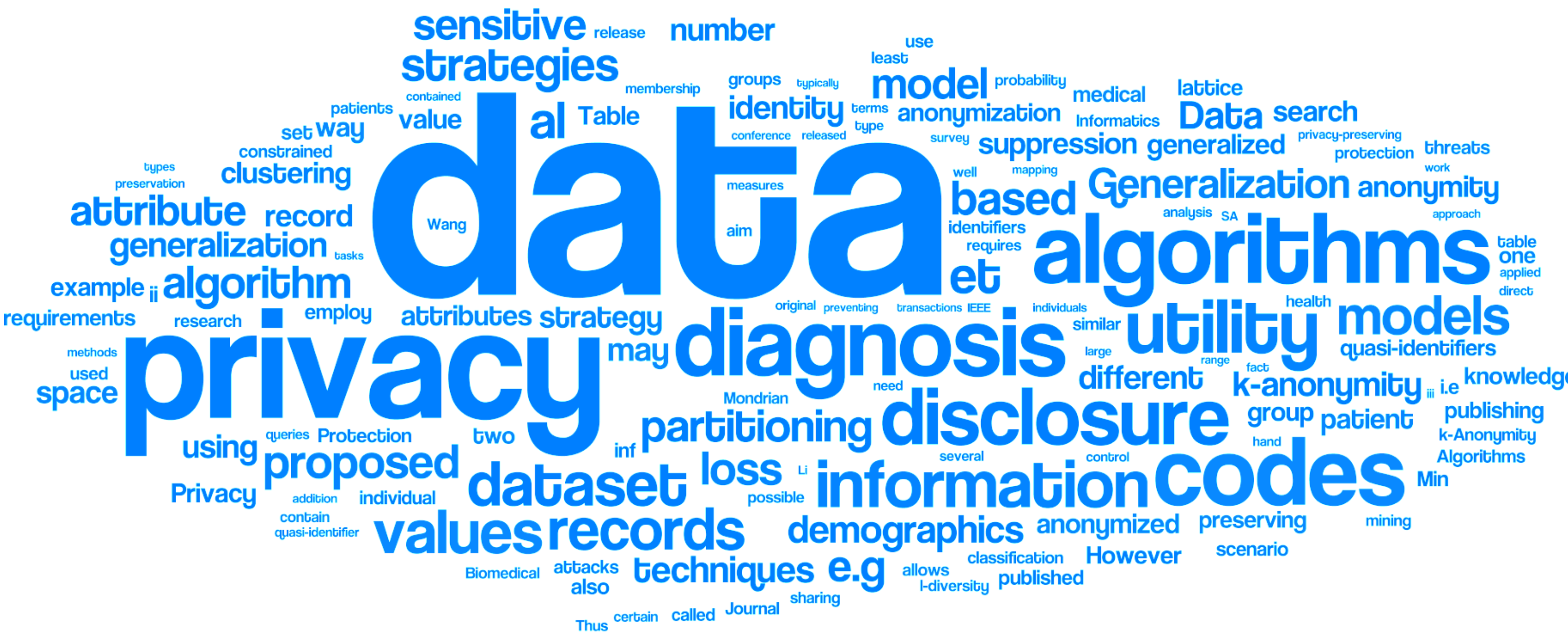
Transformation methods (cont'd)

- **Non-truthful transformations** (Pertubation) [FWF11]
  - **Post-randomization:** Randomly change categories of a categorical variable according to predefined probabilities
  - **Value distortion:** Multiplicative or additive noise
  - **Numerical rank swapping:** Randomly swap values with other values with a rank that does not differ by more than a predefined threshold
  - **Microaggregation:** Aggregate values in one group
  - **Replacing values:** Distribution sample or distribution itself
- **Methods on a structural level**
  - **Random sampling:** Randomly select a set of tuples [LQS11]
  - **Slicing:** Partition the data horizontally and vertically and creates links between partitions [LLZ+12]

Algorithms

- **Transform data to meet privacy models** [GLS14]
  - Given transformation methods, data properties etc.
- **Randomized algorithms**
  - Randomized functions for Differential Privacy [Dwork08a]
  - Genetic search [Iye02]
- **Search algorithms**
  - **Optimal algorithms:** Flash, Incognito, OLA [EDI+09, LDR05, KPE+12]
  - **Heuristic algorithms:** Top-Down-Specialization [FWY05]
- **Clustering algorithms:** Iteratively merge groups
  - **Data (focus on tuples):** Method by Tassa et al. [GT10]
  - **Space (focus on taxonomies):** For transactional data [LG13]
- **Partitioning algorithms:** Iteratively split groups
  - **Data:** Mondrian [LDR06]

Thank you for your attention!



Further Readings

- Fung CMB, Wang K, Fu A, Yu P. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques.* Chapman & Hall/CRC, ISBN: 1420091484, 2011.
- Gkoulalas-Divanis A, Loukides G, Sun J. *Publishing data from electronic health records while preserving privacy: A survey of algorithms*. J Biomed Inform, Vol. 50, p. 4-19, 2014
- Dankar FK, El Emam K. *Practicing Differential Privacy in Health Care: A Review*. Trans. Data Privacy, Vol. 6:1, 2013.
- Gkoulalas-Divanis A, Loukides G. *Anonymization of Electronic Medical Records to Support Clinical Analysis*. Springer. ISBN: 978-1-4614-5668-1, 2013
- El Emam K. *Guide to the De-Identification of Personal Health Information*. Auerbach/CRC, ISBN 978-1-4665-7906-4, 2013

References

[Agg05] Aggarwal CC. On k-anonymity and the curse of dimensionality. PVLDB, p. 901–909, 2005.
[BA05] Bayardo RJ, Agrawal R. Data Privacy through Optimal k-Anonymization. ICDE. pp. 217–228, 2005.
[BCD+07] Barak B, Chaudhuri K, Dwork C, Kale S, McSherry F et al. Privacy, accuracy, and consistency too: A holistic solution to contingency table release. PODS, p. 273–282, 2007.
[CKR+10] Cao J, Karras P, Raïssi C, Tan K. rho-uncertainty: inference-proof transaction anonymization. PVLDB;3(1):1033–44. 2010.
[Dal77] Dalenius T. Towards a methodology for statistical disclosure control. Statistisk Tidskrift, 5, p. 429–444, 1977.
[DEN+12] Dankar F, El Emam K, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. BMC Medical Informatics and Decision Making, 12:66, 2012.
[DS15] Domingo-Ferrer J, Soria-Comas J. From t-closeness to differential privacy and vice versa in data anonymization. Knowl.-Based Syst., 74:151–158, 2015.
[Dwork06] Dwork C. Differential Privacy. ICALP; 4052, pp. 1–12. Springer, 2006.
[Dwork08] Dwork C. An Ad Omnia Approach to Defining and Achieving Private Data Analysis. PinKDD, p. 1-13, 2008.
[Dwork08a] Dwork C. Differential privacy: A survey of results. TAMC. pp. 1–19, 2008.
[EDI+09] El Emam K, Dankar F, Issa R, Jonker E et al. A globally optimal k-anonymity method for the de-identification of health data. JAMIA, 16(5), pp. 670–682, 2009.
[Emam13] El Emam K. Guide to the De-Identification of Personal Health Information. Auerbach/CRC, ISBN 978-1-4665-7906-4, 2013.
[FDE13] Fida K, Dankar F, El Emam K. Practicing differential privacy in health care: A review. Trans. Data Privacy, 6(1):35–67, 2013.
[FWF11] Fung CMB, Wang K, Fu A, Yu P. Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques. Chapman & Hall/CRC, ISBN: 1420091484, 2011.
[FWY05] Fung BCM, Wang K, Yu PS. Top-Down Specialization for Information and Privacy Preservation. ICDE, 2005, pp. 205–216.
[GKK07] Ghinita G, Karras P, Kalnis P, Mamoulis N. Fast data anonymization with low information loss. VLDB, 2007, pp. 758–769.
[GT09] Gionis A, Tassa T. k-Anonymization with Minimal Loss of Information. IEEE Trans Knowl Data Eng, pp. 206–219, 2009.
[GT10] Goldberger J, Tassa T. Efficient anonymizations with enhanced utility. Transactions on data privacy, vol. 3, pp. 149–175, 2010.
[GLS14] Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. JBI, Vol. 50, p. 4-19, 2014
[HIP] U.S. Department of Health and Human Services Office for Civil Rights. HIPAA administrative simplification regulation text; 2006.
[Iye02] Iyengar VS. Transforming data to satisfy privacy constraints. SIGKDD, 2002.
[KPE+12] Kohlmayer F, Prasser F, Eckert C, Kemper A, Kuhn KA. Flash: Efficient, Stable and Optimal K-Anonymity. PASSAT., pp. 708–717, Sep. 2012.
[LDD+05] LeFevre K, DeWitt DJ, Ramakrishnan R. Multidimensional K-Anonymity (TR-1521). Tech. rep. University of Wisconsin, 2005.
[LDR05] LeFevre K, DeWitt D, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. SIGMOD. 2005, pp. 49–60, 2005.
[LDR06] LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian Multidimensional K-Anonymity. ICDE, 2006.
[LG13] Loukides G, Gkoulalas-Divanis A. Utility-aware anonymization of diagnosis codes. J Biomed Health Inform. 2013;17(1):60–70.
[LGM10] Loukides G, Gkoulalas-Divanis A, Malin B, COAT: COnstraint-based anonymization of transactions. Knowl Inf Syst, 28:2, p. 251–282, 2010.
[LLV07] Li N, Li T, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity. ICDE; p. 106–15. 2007.
[LLZ+12] Li T, Li N, Zhang J, Molloy I. Slicing: A new approach for privacy preserving data publishing. IEEE Trans Knowl Data Eng, 24(3):561–574, 2012
[LM12] Li C, Miklau G. An adaptive mechanism for accurate query answering under differential privacy. PVLDB, 5(6):514–525, 2012.
[LQS11] Li N, Qardaji WH, Su D. Provably private data anonymization: Or, k-anonymity meets differential privacy. CoRR, abs/1101.2604, 2011.
[MFH+09] Mohammed N, Fung BCM, Hung PCK, and Lee CK. Anonymizing healthcare data: A case study on the blood transfusion service. SIGKDD, p. 1285–1294, 2009.
[MGK+06] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-Diversity: privacy beyond k-anonymity. ICDE; pp. 24, 2006.
[NAC07] Nergiz ME, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. SIGMOD; p. 665–676, 2007.
[NC10] Nergiz ME, Clifton C. d-presence without complete world knowledge. IEEE Trans Knowl Data Eng; 22(6):868–83, 2010.
[PK08] Prasad S, Kasiviswanathan AS. A note on differential privacy: Defining resistance to arbitrary side information. CoRR abs/0803.3946, 2008.
[Swe02] Sweeney L. k-anonymity: a model for protecting privacy. IJUFKS;10:557–70, 2002.
[TMK08] Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set valued data. PVLDB;1(1):115–25, 2008.
[TMK08] Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of setvalued data. PVLDB;1(1):115–25, 2008.
[TV06] Truta TM, Vinay B. Privacy protection: p-sensitive k-anonymity property. ICDE workshops; p. 94, 2006.
[VI02] Iyengar VS. Transforming data to satisfy privacy constraints. SIGKDD, pp. 279–288, 2002.
[XT07] Xiao X, Tao Y. m-invariance: Towards privacy preserving republication of dynamic datasets. SIGMOD, 2007.
[XWF+08] Xu Y, Wang K, Fu AW-C, Yu PS. Anonymizing transaction databases for publication. KDD; p. 767–75, 2008.