Partition-Based Models and Algorithms for Privacy-Preserving Data Publishing

M2 SIF - SED

Tristan Allard
Univ. Rennes 1 / Irisa lab.
tristan.allard@irisa.fr

Autumn 2023

Targeted Ad'

Looking for an internship (or more ?)

- Engineering, **research**, both
- Protection of individuals' data : privacy, explainability, fairness
- **System** orientation (not pure theory)

Progress of the Talk

Progress of the Talk

k-Anonymity : Assumptions I

- Considers that individuals' data is made of :
  - Identifying attributes, or **ID**: **identify uniquely** each individual (e.g., $\langle$`SSN`$\rangle$);
  - Quasi-Identifying attributes, or **QID**: **may identify uniquely** some individuals (*e.g.*, $\langle$ `Zip`, `DoB`$\rangle$);
  - Sensitive attributes, or **SD**: sensitive data, e.g., $\langle$ `Disease` $\rangle$;

k-Anonymity : Assumptions II

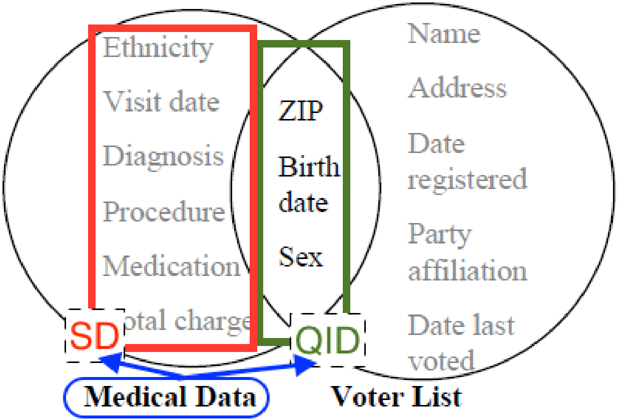


Figure: Quasi-identifiers and sensitive data in Gov. Weld's case

k-Anonymity: the Model I

Warning

We consider in this talk that each individual has a single record in the DB.

k-Anonymity: the Model II

A release is k-anonymous [14, 16] if:

- It does not contain any direct identifier
- The QID of each record has been made indistinguishable from at least $(k-1)$ others

$\Rightarrow$ Each sensitive data is within a group that corresponds to at least k QID.

k-Anonymity: the Model III

Name	Zip	Age	Dis.
Bob	75001	22	Cold
Bill	75002	29	Flu
Don	75003	22	Cold
Sue	75010	28	HIV

Table: Raw data (*e.g.*, GIC medical data).

Zip	Age	Dis.
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Table: A possible 2-Anonymous Release of the raw data.

k-Anonymity: the Model IV

Name	Zip	Age
Bob	75001	22

Zip	Age	Dis.
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Table: Left: External knowledge made of a known QID (*e.g.*, voter list). Right: A possible 2-Anonymous release of the raw data.

$\Rightarrow$ Joins on QID are now ambiguous: what is Bob's disease?

k-Anonymity: the Model V

Vocabulary

- **Equivalence class**: A group of records indistinguishable wrt their QID
- **Sanitized release**: the set of equivalence classes finally published

Progress of the Talk

- k-Anonymity VS Pseudonymity : "Hide into the crowd"
  - Formal Model
  - Algorithms for k-Anonymity
  - Deducing the Three Components of a PPDP Solution
- l-Diversity : "Ensure the Crowd is Diverse Enough"
- Endless Cycle
- Conclusion
- References

Achieving *k*-Anonymity

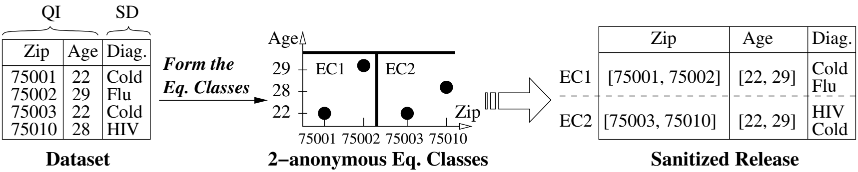
- **Generalization of the quasi-identifiers** : the most used operation (the more general a value is, the more people correspond to it : "less people in Urrugne, than in Pays Basque, than in France.")
  - Numerical attribute : from values to ranges
  - Categorical attribute : need a taxonomy (*e.g.*, Urrugne > Pays Basque > France)
- **Optimality is too hard** : not all generalizations are equal (the less you generalize the more accurate the data).
  $\Rightarrow$ How to output a release that satisfies *k*-Anonymity with a minimal *number of generalizations* ? Shown to be hard [1, 13]
- **Many alternative strategies/simplifications/heuristics exist** (e.g., [1, 2, 5, 7, 15, 13, 17])

Not the focus of this talk but lets have a quick look at one of them. . .

Mondrian : A Simple Algorithm for Achieving *k*-Anonymity

- **Goal**: form equivalence classes that span at least k similar QID values
- **How?** Greedily !
  - Starts with one *partition* of the dataset containing all the records
  - Recursively partitions it into smaller and smaller partitions
  - Finally replace the QID value of each record by the range of its partition

Mondrian Illustrated



In this example, we want 2-Anonymity (at least two records per class).

MONDRIAN in details I

Algorithm 1: `MondrianAnonymize`

input : A partition $\mathcal{P}$ to split

output: A set of partitions, each containing between k and $2k - 1$ tuples

if *no allowable multidimensional cut for partition* **then return** $\mathcal{P}$;

else

  $dim \leftarrow$ `chooseDimension()`;

  $fs \leftarrow$ `frequencySet(`$\mathcal{P}$, dim`)`;

  $splitVal \leftarrow$ `findMedian(`fs`)`;

  $\mathcal{L} \leftarrow \{t \in \mathcal{P} : t.dim \leq splitVal\}$;

  $\mathcal{R} \leftarrow \{t \in \mathcal{P} : t.dim > splitVal\}$;

  return `MondrianAnonymize(`$\mathcal{L}$`)` $\cup$ `MondrianAnonymize(`$\mathcal{R}$`)`

Mondrian in details II

`MondrianAnonymize` internal calls:

- `chooseDimension`: choose the dimension in which to split (usually the widest one);
- `frequencySet`: set of unique values taken by the tuples for the chosen dimension, each paired with the number of times it appears;
- `findMedian`: find the median;

Mondrian, for Real I

Actually, Mr Mondrian was a painter !

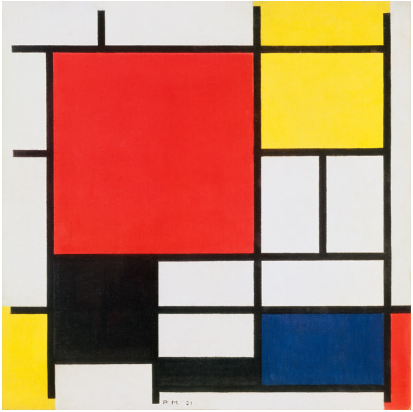


Figure: Composition en rouge, jaune, bleu et noir. Mondrian. 1926

Mondrian, for Real II

And a `MondrianAnonymize` partitioning may look like this :

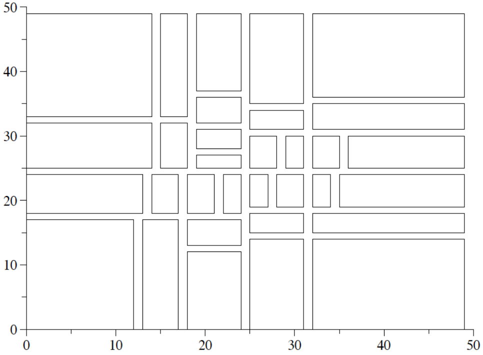


Figure: Example of a Mondrian partitioning [8] (synthetic data, 1000 tuples, k=25, normal distribution).

Progress of the Talk

- k-Anonymity VS Pseudonymity : “Hide into the crowd”
  - Formal Model
  - Algorithms for k-Anonymity
  - Deducing the Three Components of a PPDP Solution
- l-Diversity : “Ensure the Crowd is Diverse Enough”
- Endless Cycle
- Conclusion
- References

Synthesis

How do you position the elements we just saw with respect to the usual components of a PPDP solution ?

Progress of the Talk

Progress of the Talk

Some Defects of k-Anonymity

Name	Zip	Age
Bob	75001	22

Zip	Age	Dis.
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Table: Attack considered by k-Anonymity. Left: External knowledge made of a known QID (*e.g.*, voter list). Right: A possible 2-Anonymous release.

1. **Homogeneity**: What if all the SD of the QI of an equivalence class are identical?
2. **Background knowledge**: What if the adversary knows that his victim is more or less likely to have a given sensitive data?

$\Rightarrow$ Motivate the l-Diversity model

Foundation: the Bayes-Optimal Privacy Model I

Founding intuition

Background knowledge about SD should be **expressed** and **taken into account** by the privacy model.

The Bayes-Optimal Privacy model [11] is an early attempt to this end (2006):

- **Background knowledge**: joint distribution between QI and SD
- **Prior belief**: given a targeted QI q and a SD s, probability of s given q
- **Posterior belief**: given a targeted QI q, a SD s, **and the sanitized release** $\mathcal{V}$, probability of s given q and $\mathcal{V}$
- **Privacy breach**: if $\textit{distance}(\textit{posterior belief}, \textit{prior belief}) > \theta$ (too much gain in knowledge)

Foundation: the Bayes-Optimal Privacy Model II

The intuition behind THIS definition of a privacy breach is **a way to envision privacy** (also called a *paradigm* in these slides) !

Paradigm#1: **Uninformative Principle** [11]

A privacy breach occurs when the *prior belief* of the adversary differs *significantly* from his *posterior belief*.

"If the **release of the statistics S** *make it possible to determine the value D_k* **more accurately** *than is possible* **without access to S**, *disclosure has taken place (. . .)"*
Dalenius 1977 [4]

Formalizing the Bayes-Optimal Model I

- Background knowledge: joint distribution between quasi-identifiers and sensitive data : $f(s, q)$.

Prior belief

Given a target QI q (the victim) and a sensivite data s :

$$\alpha(q, s) = \mathtt{Pr}_f(s|q) = \frac{f(s, q)}{\sum_{s' \in SD} f(s', q)} \tag{1}$$

Formalizing the Bayes-Optimal Model II

- Let $\mathcal{V}$ be the sanitized release
- Let $q^\star$ be the QI of the equivalence class that contains q
- Let $n(q^\star, s)$ be the number of tuples $\langle q^\star, s\rangle$ in $\mathcal{V}$;
- Let $f(s|q^\star)$ be the conditional probability that s be associated to the QIs that have been generalized to $q^\star$;

Posterior belief

Given a target QI q, a sensitive data s, and the release $\mathcal{V}$:

$$\beta(q, s, \mathcal{V}) = \mathtt{Pr}(s|q \wedge \mathcal{V}) = \frac{n(q^\star, s)\frac{f(s|q)}{f(s|q^\star)}}{\sum_{s' \in SD} n(q^\star, s')\frac{f(s'|q)}{f(s'|q^\star)}} \tag{2}$$

(proof in [11])

Formalizing the Bayes-Optimal Model III

A sanitized release $\mathcal{V}$ satisfies BAYES-OPTIMAL PRIVACY if:

$$\forall q \in QI, s \in SD, \texttt{abs}(\alpha(q,s) - \beta(q,s,\mathcal{V})) < \theta \tag{3}$$

where abs returns the absolute value of its argument and θ is the user-defined threshold over the adversarial knowledge gain.
Note: alternative definitions exist [11].

Example I

Let the adversary's background knowledge about Don be:

$f(\langle\ q_{Don}, Cold\rangle) = 0.1$	$\alpha(q_{Don}, Cold) = ??$
$f(\langle q_{Don}, Flu\rangle) = 0.01$	$\alpha(q_{Don}, Flu) = ??$
$f(\langle q_{Don}, HIV\rangle) = 0.14$	$\alpha(q_{Don}, HIV) = ??$

What is his prior belief about Don ?

Example II

Answer:

$f(\langle\ q_{Don}, Cold\rangle) = 0.1$	$\alpha(q_{Don}, Cold) = 0.1/0.25 = 0.4$
$f(\langle q_{Don}, Flu\rangle) = 0.01$	$\alpha(q_{Don}, Flu) = 0.01/0.25 = 0.04$
$f(\langle q_{Don}, HIV\rangle) = 0.14$	$\alpha(q_{Don}, HIV) = 0.14/0.25 = 0.56$

Example III

Let the adversary's background knowledge about any individual other than Don be:

$f(\langle\ q_i, Cold\rangle) = 0.083$	$\alpha(q_i, Cold) = ??$
$f(\langle q_i, Flu\rangle) = 0.083$	$\alpha(q_i, Flu) = ??$
$f(\langle q_i, HIV\rangle) = 0.083$	$\alpha(q_i, HIV) = ??$

What is his prior belief about any other individual ?

Example IV

Answer:

$f(\langle\ q_i, Cold\rangle) = 0.083$	$\alpha(q_i, Cold) = 0.083/0.25 = 0.33$
$f(\langle q_i, Flu\rangle) = 0.083$	$\alpha(q_i, Flu) = 0.083/0.25 = 0.33$
$f(\langle q_i, HIV\rangle) = 0.083$	$\alpha(q_i, HIV) = 0.083/0.25 = 0.33$

Example V

Let $\mathcal{V}$ be the 2-anonymous release:

Zip	**Age**	**Dis.**
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Recall that $q_{Don} = \langle 75003, 22 \rangle$ and is known by the adversary.

What is his posterior belief about Don ?

Example VI

Answer:

In the above release, $q^{\star}_{Don} = \langle[75003, 75010], [22, 29]\rangle$.

Then, the adversary's posterior belief about Don is:

$$\boxed{\begin{array}{l} \beta(q_{Don}, Flu, \mathcal{V}) = \frac{0 * \frac{0.04}{0.37}}{1.18} = 0 \\ \beta(q_{Don}, Cold, \mathcal{V}) = \frac{1 * \frac{0.4}{0.73}}{1.18} = 0.46 \\ \beta(q_{Don}, HIV, \mathcal{V}) = \frac{1 * \frac{0.56}{0.89}}{1.18} = 0.54 \end{array}}$$

Example VII

As a result:

Prior	**Posterior**
$\alpha(q_{Don}, Cold) = 0.4$	$\beta(q_{Don}, Cold, \mathcal{V}) = 0.46$
$\alpha(q_{Don}, Flu) = 0.04$	$\beta(q_{Don}, Flu, \mathcal{V}) = 0$
$\alpha(q_{Don}, HIV) = 0.56$	$\beta(q_{Don}, HIV, \mathcal{V}) = 0.54$

Is there a privacy breach ?

Bayes-Optimal Privacy : Impractical

If Bayes-Optimal Privacy were practical, it could permit to check that releases do not allow significant knowledge gains...

But :

- Obtaining the joint distribution f that represents the adversarial background knowledge ?
- What if there are several adversaries ?
- What about other kinds of knowledge ?
- Cost of checking all the possible (q, s) pair !

l-Diversity I

l-Diversity: a simple and easy-to-check condition for protecting against **SD homogeneity** and **adversarial negation statements**.

l-Diversity II

l-Diversity

An l-diverse equivalence class contains at least l *well-represented* sensitive values.

l-Diversity III

"Well-represented" can be instantiated in many ways, among which:

- Naive l-Diversity : at least l distinct values appear ;
- Entropy l-Diversity: the entropy of the set of SD in each equivalence class should be at least $\log l$;
- Recursive (c, l)-Diversity: if the most frequent SD in a class is not much more frequent than the other SD of the class
- `(Put your idea here)`-Diversity

Recursive (c, l)-Diversity

For each class:

- Count the occurence of each sensitive value;
- and sort them by descending order.

Let r_1 be the first count, ..., r_m be the m^{th}.

Recursive (c, l) Diversity

An equivalence class satisfying Recursive (c, l)-Diversity satisfies: $r_1 < c(r_l + r_{l+1} + ... + r_m)$.
A release $\mathcal{V}$ satisfies Recursive (c, l)-Diversity if all its equivalence classes satisfy it.

Examples

What is the protection offered by the classes having the following counts?

r_1	100
r_2	6
r_3	5
r_4	3

Examples

What is the protection offered by the classes having the following counts?

r_1	100
r_2	6
r_3	5
r_4	3

r_1	7
r_2	6
r_3	5
r_4	3

Recursive (c, l) Diversity, bis I

Assume that the counts of Don's class are as follows:

r_1	7
r_2	6
$\boxed{r_3}$	5
r_4	3
r_5	1
r_6	1

$\Rightarrow$ Satisfies RECURSIVE $(1,\ 3)$-DIVERSITY.

Recursive (c, l) Diversity, bis II

The adversary knows that Don **does not** have flu.

If the count of flu is r_2:

$$\begin{array}{|c|c|} r_1 & 7 \\ r_2 & 6 \\ \boxed{r_3} & 5 \\ r_4 & 3 \\ r_5 & 1 \\ r_6 & 1 \end{array} \Rightarrow \begin{array}{|c|c|} r_1 & 7 \\ \boxed{r_2} & 5 \\ r_3 & 3 \\ r_4 & 1 \\ r_5 & 1 \end{array}$$

$\Rightarrow$ Satisfies Recursive $(1,\ 2)$-Diversity.

Recursive (c, l) Diversity, bis III

The adversary knows that Don **does not** have flu.

If the count of flu is r_6:

$$\begin{array}{|c|c|} r_1 & 7 \\ r_2 & 6 \\ \boxed{r_3} & 5 \\ r_4 & 3 \\ r_5 & 1 \\ r_6 & 1 \end{array} \Rightarrow \begin{array}{|c|c|} r_1 & 7 \\ r_2 & 6 \\ \boxed{r_3} & 5 \\ r_4 & 3 \\ r_5 & 1 \end{array}$$

$\Rightarrow$ Satisfies Recursive $(1,\ 3)$-Diversity.

Recursive (c, l) Diversity, bis IV

Recursive (c, l)-Diversity $+$ 1 negation statement $\rightarrow$ What is the protection level at worst?

Progress of the Talk

k-Anonymity VS Pseudonymity : “Hide into the crowd”

l-Diversity : “Ensure the Crowd is Diverse Enough”

- The Model(s)
- The Algorithms

Endless Cycle

Conclusion

References

Simple Updates to k-Anonymity Algorithms

- Use algorithms designed for achieving k-Anonymity
- Add as an additional constraint on the equivalence classes the l-Diversity criterion

Progress of the Talk

The Family of Partition-Based Models and Algorithms

Many followers, based on producing equivalence classes by generalizing the QID.

Gave rise to the family of partition-based approaches :

1. Remove the ID attribute(s)
2. Form groups of records (partitions) according to the values of QID and SD of the actual records
3. And finally disclose statistical information (really !) at the group level.

Weaknesses

- Proposal (year n) $\rightarrow$ Attack or limit + fix (year $n+1$)
- Various severe attacks/limits exist:
  - **No composability**: intersecting the respective sets of QID and of SD of two non-disjoint k-Anonymous releases may break k-Anonymity [19]
  - **Leaks in the execution sequences** (for optimality) : execution sequence depends on data $\Rightarrow$ minimality attacks [18]
  - **Naive adversarial reasonning models** : adversarial correlations between the QID and SD values of an equivalence class ignore the other classes $\Rightarrow$ Model the correlations between QID and SD values, in all the classes, by a bayesian network with probabilistic parameters (*aka* deFinetti attacks) [6]
  - **Numerous possible types of background knowledge** : negation statements [11], distribution of SD in the dataset [9], joint distribution between QID and SD [10, 11], logical sentences [3, 12], etc.

$\Rightarrow$ Is pursuing this cycle worth ?

RIP Partition-Based Approaches ?

Today :

- Partition-based approaches have been shown to suffer from many flaws
- Strong interest decrease from academics
- *Differential privacy* and models inspired from it take the lead (see next lecture)
- But...
  "Nous sommes en 50 avant Jésus-Christ. Toute la Gaule est occupée par les Romains... Toute ? Non ! Car un village peuplé d'irréductibles Gaulois résiste encore et toujours à l'envahisseur."
  - Some real-world organizations are enclined to use it (intuitive models, illusion of retaining "true data")
  - European "CNILs" (*i.e.*, the Article 29 Data Protection Working Party) refer to these models as possible approaches for sanitizing data [1]

[1] See the Opinion 05/2014 on Anonymisation Techniques `ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf`

Progress of the Talk

Conclusion

- Advantages : partition-based models are intuitive
- Drawbacks : assumes that attributes can be partitioned across QID/SD, fixes needed for breaches (minimality attacks and others) and for composition issues, algorithms are costly, illusion of utility ("true records are disclosed"), *etc.*
- Must be known : you may encounter them (well-known models) but think twice before using them !

Progress of the Talk

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu.
Anonymizing tables.
In *Proceedings of the 10th International Conference on Database Theory*, ICDT'05, pages 246–258, Berlin, Heidelberg, 2005. Springer-Verlag.

[2] R. J. Bayardo and R. Agrawal.
Data privacy through optimal k-anonymization.
In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, pages 217–228, Washington, DC, USA, 2005. IEEE Computer Society.

[3] B.-C. Chen, K. LeFevre, and R. Ramakrishnan.
Privacy skyline: privacy with multidimensional adversarial knowledge.
In *Proceedings of the 33rd international conference on Very large data bases*, VLDB '07, pages 770–781. VLDB Endowment, 2007.

[4] T. Dalenius.

Towards a methodology for statistical disclosure control.
Statistik Tidskrift, 15(5):429–444, 1977.

[5] B. C. M. Fung, K. Wang, and P. S. Yu.
Top-down specialization for information and privacy preservation.
In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, pages 205–216, Washington, DC, USA, 2005. IEEE Computer Society.

[6] D. Kifer.
Attacks on privacy and deFinetti's theorem.
In *Proceedings of the 35th SIGMOD international conference on Management of data*, SIGMOD '09, pages 127–138, New York, NY, USA, 2009. ACM.

[7] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan.
Incognito: Efficient full-domain k-anonymity.
In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 49–60, New York, NY, USA, 2005. ACM.

[8] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan.
Mondrian multidimensional k-anonymity.
In *Proceedings of the 22nd International Conference on Data Engineering*, ICDE '06, pages 25–, Washington, DC, USA, 2006. IEEE Computer Society.

[9] N. Li, T. Li, and S. Venkatasubramanian.
t-closeness: Privacy beyond k-anonymity and l-diversity.
In *Proceedings of the 23rd IEEE International Conference on Data Engineering*, ICDE '07, pages 106–115, april 2007.

[10] A. Machanavajjhala, J. Gehrke, and M. Götz.
Data publishing against realistic adversaries.
PVLDB, 2(1):790–801, August 2009.

[11] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam.
ℓ-diversity: Privacy beyond κ-anonymity.
In *Proceedings of the 22nd IEEE International Conference on Data Engineering*, ICDE '06, pages 24–, Washington, DC, USA, 2006. IEEE Computer Society.

[12] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern.
Worst-case background knowledge for privacy-preserving data publishing.
In *Proceedings of the 23rd IEEE International Conference on Data Engineering*, pages 126–135, 2007.

[13] A. Meyerson and R. Williams.
On the complexity of optimal k-anonymity.
In *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, pages 223–228, New York, NY, USA, 2004. ACM.

[14] P. Samarati.
Protecting respondents identities in microdata release.
IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027, 2001.

[15] P. Samarati and L. Sweeney.

Generalizing data to provide anonymity when disclosing information (abstract).
In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, PODS '98, pages 188–, New York, NY, USA, 1998. ACM.

[16] L. Sweeney.
k-anonymity: a model for protecting privacy.
Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5):557–570, 2002.

[17] K. Wang, P. S. Yu, and S. Chakraborty.
Bottom-up generalization: A data mining solution to privacy protection.
In *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04, pages 249–256, Washington, DC, USA, 2004. IEEE Computer Society.

[18] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei.
Minimality attack in privacy preserving data publishing.

In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, pages 543–554. VLDB Endowment, 2007.

[19] X. Xiao and Y. Tao.
M-invariance: Towards privacy preserving re-publication of dynamic datasets.
In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 689–700, New York, NY, USA, 2007. ACM.