IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-6 E-ISSN: 2347-2693

# m-Privacy Preserving Data Analysis And Data Publising

Sanjeev Rathod[1*] and Doddegowda.B.J [2]

[1] *M.Tech (Software Engineering), VTU University, INDIA*
[2] *Computer Science And Engineering, VTU University, INDIA*

rathod.sanjeev@gmail.com[1*]; bjdgowda10@gmail.com[2]

www.ijcaonline.org

Received: 21 May 2014 Revised: 07 Jun 2014 Accepted: 18 Jun 2014 Published: 30 Jun 2014

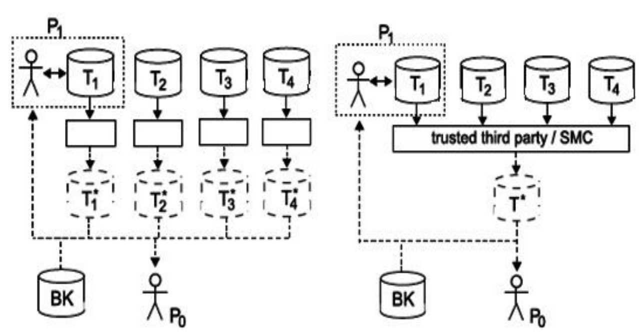
***Abstract—*** Combining and analyzing data collected at multiple administrative locations is critical for a wide variety of applications, such as detecting malicious attacks or computing an accurate estimate of the popularity of Web sites. However, legitimate concerns about privacy often inhibit participation in collaborative data analysis. In this paper, we design, implement, and evaluate a practical solution for privacy-preserving data analysis and data publishing among a large number of participants. There is an increasing need for sharing data that contain personal information from distributed databases. For example, in the healthcare domain, a national agenda is to develop the Nationwide Health Information Network (NHIN) to share information among hospitals and other providers, and support appropriate use of health information beyond direct patient care with privacy protection. Privacy preserving data analysis and data publishing has received considerable attention in recent years as promising approaches for sharing data while preserving individual privacy. When the data are distributed among multiple data providers or data owners, two main settings are used for anonymization. One approach is for each provider to anonymize the data independently (anonymize-and-aggregate), which results in potential loss of integrated data utility.

***Keywords—*** m-Privacy, k-anonymity, l-diversity, Database Management, Heuristic algorithms, Distributed Data Publising, Pruning Strategies.

## I. INTRODUCTION

Data mining is the process of extracting useful, interesting, and previously unknown information from large data sets. The success of data mining relies on the availability of high quality data and effective information sharing. The collaborative data publishing setting (Figure 1b) with horizontally partitioned data across multiple data providers, each contributing a subset of records Ti. As a special case, a data provider could be the data owner itself who is contributing its own records. This is a very common scenario in social networking and recommendation systems. A more desirable approach is collaborative data publishing, which anonymizes data from all providers as if they would come from one source (aggregate and- anonymize), using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols to do computations. Considering different types of malicious users and information they can use in attacks, we identify three main categories of attack scenarios. While the first two are addressed in existing work, the last one receives little attention and will be the focus of this paper. Considering different types of malicious users and information they can use in attacks, we identify three main categories of attack scenarios. While the first two are addressed in existing work, the last one receives little attention and will be the focus of this paper. A task of the utmost importance is to develop methods and tools for publishing data in a hostile environment so that the published data remain practically useful while individual privacy is preserved. This undertaking is called privacy-preserving data publishing (PPDP), which can be viewed as a technical response to complement the privacy policies categories of attack scenarios. While the first two are addressed in existing work, the last one receives little attention and will be the focus of this paper. A task of the utmost importance is to develop methods and tools for publishing data in a hostile environment so that the published data remain practically useful while individual privacy is preserved. This undertaking is called privacy-preserving data publishing (PPDP), which can be viewed as a technical response to complement the privacy policies.

A B




(a) Anonymize and aggregate (b) Aggregate and anonymize

Corresponding Author: Mr. *SANJEEV, rathod.sanjeev@gmail.com*

Fig. 1 Distributed Data Publishing Settings for Four Providers

## II. RELATED WORK

Privacy preserving data analysis and data publishing has received considerable attention in recent years as promising approaches for sharing data while preserving individual privacy. In a non-interactive model, a data provider publishes a "sanitized" version of the data, simultaneously providing utility for data users, and privacy protection for the individuals represented in the data. When data are gathered from multiple data providers or data owners, two main settings are used for anonymization. One approach is for each provider to anonymize the data independently, which results in potential loss of integrated data utility.

Privacy preserving data analysis and publishing has received considerable attention in recent years. Most work has focused on a single data provider setting and considered the data recipient as an attacker. A large body of literature assumes limited background knowledge of the attacker, and defines privacy using relaxed *adversarial* notion by considering specific types of attacks. Representative principles include *k*-anonymity, *l*-diversity, and *t*-closeness. A few recent works have modeled the instance level background knowledge as corruption, and studied perturbation techniques under these syntactic privacy notions. In the distributed setting that we study, since each data holder knows its own records, the *corruption* of records is an inherent element in our attack model, and is further complicated by the collusive power of the data providers. On the other hand, differential privacy is an unconditional privacy guarantee but only for statistical data release or data computations.

Collaborative data publishing can be considered as a multi-party computation problem, in which multiple providers wish to compute an anonymized view of their data without disclosing any private and sensitive information. We assume the data providers are semi-honest, commonly used in distributed computation setting. A trusted third party (TTP) or Secure Multi-Party Computation (SMC) protocols can be used to guarantee there is no disclosure of intermediate information during the anonymization. However, neither TTP nor SMC protects against inferring information using the anonymized data.

*Disadvantages,* Malicious users are colluding the data (related to shilling attackers).

Anonymization techniques are not control the all different attackers.

## III. SYSTEM DESIGN

### A. Patient Registration

In this module if a patient has to take treatment, he/she should register their details like Name, Age, and Disease they get affected, Email etc.These details are maintained in a Database by the Hospital management. Only Doctors can see all their details. Patient can only see his own record.

Based on this Paper, When the data are distributed among multiple data providers or data owners, two main settings are used for anonymization. One approach is for each provider to anonymize the data independently (anonymize-and-aggregate, Figure 1A), which results in potential loss of integrated data utility. A more desirable approach is collaborative data publishing which anonymizes data from all providers as if they would come from one source (aggregate-and-anonymize, Figure 1B), using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols to do computations

### B. Attacks by External Data Recipient Using Anonymized Data

A data recipient, e.g. P0, could be an attacker and attempts to infer additional information about the records using the published data (T∗) and some background knowledge (BK) such as publicly available external data.

### C. Attacks by Data Providers Using Anonymized Data and Their Own Data

Each data provider, such as P1 in Figure 1, can also use anonymized data T∗ and his own data (T1) to infer additional information about other records. Compared to the attack by the external recipient in the first attack scenario, each provider has additional data knowledge of their own records, which can help with the attack. This issue can be further worsened when multiple data providers collude with each other.

### D. Doctor Login

In this module Doctor can see all the patients details and will get the background knowledge(BK),by the chance he will see horizontally partitioned data of distributed data base of the group of hospitals and can see how many patients are affected without knowing of individual records of the patients and sensitive information about the individuals.

### E. Admin Login

In this module Admin acts as Trusted Third Party (TTP).He can see all individual records and their sensitive information among the overall hospital distributed data base. Anonymation can be done by this people. He/She collected information's from various hospitals and grouped into each other and makes them as an anonymised data.

We illustrate the *m*-adversary threats with an example shown in Table I. Assume that hospitals *P*1, *P*2, *P*3, and *P*4 wish to collaboratively anonymize their respective patient databases *T*1, *T*2, *T*3, and *T*4. In each database, Name is an identifier, {Age, Zip} is a quasi-identifier (QI), and Disease is a sensitive attribute. *T*a* is one possible QI-group-based anonymization using existing approaches that guarantees *k*-anonymity and *l*-diversity ($k = 3$, $l = 2$). Note that *l*-diversity holds if each equivalence group contains records with at least *l* different sensitive attribute values. However, a tacker from the hospital *P*1, who has access to *T*1, may remove all records from *T*a* is also in *T*1 and find out that there is only one patient between 20 and 30 years old. Combining this information with background knowledge *BK*, *P*1 can identify Sara's record (highlighted in the table) and her disease Epilepsy. In general, multiple providers may collude with each other, hence having access to the union of their data, or a user may have access to multiple databases, e.g. a physician switching to another hospital, and use the increased data knowledge to infer data at other nodes.

$T_a^*$

| Provider | Name | Age | Zip | Disease |
|---|---|---|---|---|
| $P_1$ | Alice | [20-30] | ***** | Cancer |
| $P_1$ | Emily | [20-30] | ***** | Asthma |
| $P_3$ | **Sara** | **[20-30]** | ***** | **Epilepsy** |
| $P_1$ | Bob | [31-35] | ***** | Asthma |
| $P_2$ | John | [31-35] | ***** | Flu |
| $P_4$ | Olga | [31-35] | ***** | Cancer |
| $P_4$ | Frank | [31-35] | ***** | Asthma |
| $P_2$ | Dorothy | [36-40] | ***** | Cancer |
| $P_2$ | Mark | [36-40] | ***** | Flu |
| $P_3$ | Cecilia | [36-40] | ***** | Flu |

$T_b^*$

| Provider | Name | Age | Zip | Disease |
|---|---|---|---|---|
| $P_1$ | Alice | [20-40] | ***** | Cancer |
| $P_2$ | Mark | [20-40] | ***** | Flu |
| $P_3$ | Sara | [20-40] | ***** | Epilepsy |
| $P_1$ | Emily | [20-40] | 987** | Asthma |
| $P_2$ | Dorothy | [20-40] | 987** | Cancer |
| $P_3$ | Cecilia | [20-40] | 987** | Flu |
| $P_1$ | Bob | [20-40] | 123** | Asthma |
| $P_4$ | Olga | [20-40] | 123** | Cancer |
| $P_4$ | Frank | [20-40] | 123** | Asthma |
| $P_2$ | John | [20-40] | 123** | Flu |

Table.1 m-privacy and m-adversary

## IV. DEFINITION OF M-PRIVACY

m-privacy definition with respect to a given privacy constraint to prevent inference attacks by *m*-adversary, followed by its properties.

Let T = {t1, t2, . . .}be a set of records horizontally distributed among *n* data providers $P = \{P_1, P_2, ..., P_n\}$,such that $T_i \subseteq T$ is a set of records provided by $P_i$.We assume $A_S$ is a sensitive attribute with domain $D_S$. If the records contain multiple sensitive attributes then a new sensitive attribute $A_S$ can be defined as a Cartesian product of all sensitive attributes. Our goal is to publish an anonymized table $T^*$ while preventing any *m*-adversary from inferring $A_S$ for any single record.

*A. m-Privacy*

To protect data from external recipients with certain background knowledge *(BK)* we assume a given privacy requirement *C*, defined by a conjunction of privacy constraints:

$C_1 \wedge C_2 \wedge ... \wedge C_w$. If a set of records $T^*$ satisfies *C*, we say $C(T^*) = true$. Any of the existing privacy principles can be used as a component constraint.

In our example (Table I), the privacy constraint *C* is defined as $C = C_1 \wedge C_2$, where $C_1$ is *k*-anonymity with $k = 3$, and $C_2$ is *l*-diversity with $l = 2$. Both anonymized tables, $T_a^*$ and $T_b^*$ satisfies *C*, although as we have shown earlier, $T^*$ may be compromised by an *m*-adversary such as $P_1$.

We now formally define a notion of *m*-privacy with respect to a privacy constraint *C*, to protect the anonymized data against *m*-adversaries in addition to the external data recipients. The notion explicitly models the inherent data knowledge of an *m*-adversary, the data records they jointly contribute, and requires that each equivalence group, excluding any of those records owned by an *m*-adversary, still satisfies *C*.

**Definition:** *m*-PRIVACY Given *n* data providers, a set of records *T*, and an anonymization mechanism *A*, an *m*-adversary *I* $(m \leq n-1)$ is a coalition of *m* providers, which jointly contributes a set of records $T_I$.

Sanitized records $T^* = A(T)$ satisfy *m*-privacy, i.e. are *m*-private, with respect to a privacy constraint *C*, if and only if, $\forall I \subset P, |I| = m, \forall T' \subseteq T : T' \supseteq T \setminus T_I, C(A(T')) = true$.

## V. M-PRIVACY VERIFICATION

Checking whether a set of records satisfies m-privacy creates a potential computational challenge due to the combinatorial number of m-adversaries that need to be checked. In this section, we first analyze the problem by modeling the checking space. Then we present heuristic algorithms with effective pruning strategies and adaptive ordering techniques for efficiently checking m-privacy for a set of records w.r.t. an EG monotonic privacy constraint C.

### i. Adversary Space Enumeration

Given a set of nG data providers, the entire space of m-adversaries (m varying from 0 to nG - 1) can be represented using a lattice shown in Figure 2. Each node at layer m represents an m-adversary of a particular combination of m providers. The number of all possible m-adversaries is equal to (nG m) Each node has parents (children) representing their data direct super- (sub-) coalitions. For simplicity the space is also represented as a diamond, where a horizontal line corresponds to all m-adversaries with the same m value, the bottom node corresponds to 0-adversary (external data recipient), and the top line to (nG-1) adversaries.

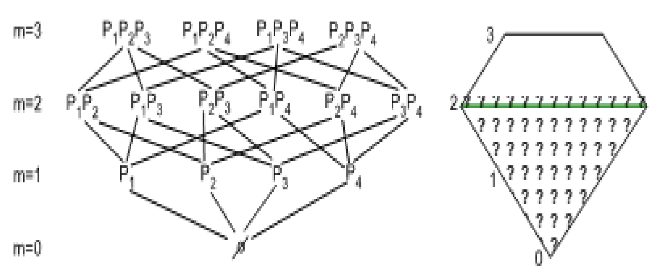



Fig. 2 m-Adversary space

### ii. Heuristic Algorithms

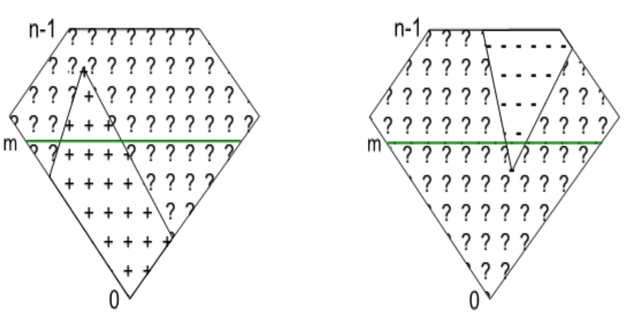
The key idea of our heuristic algorithms is to efficiently search through the adversary space with effective pruning such that not all m-adversaries need to be checked. This is achieved by two different pruning strategies, an adversary ordering technique, and a set of search strategies that enable fast pruning.

**Pruning Strategies,** The pruning strategies are possible thanks to the EG monotonicity of m-privacy. If a coalition is not able to breach privacy, then all its sub-coalitions will not be able to do so and hence do not need to be checked (downward pruning). On the other hand, if a coalition is able to breach privacy, then all its super-coalitions will be able to do so and hence do not need to be checked
(upward pruning). In fact, if a sub-coalition of an m-adversary is able to breach privacy, then the upward pruning allows the algorithm to terminate immediately as the m-adversary will be able to breach privacy (early stop). Figure 3 illustrates the two pruning strategies where + represents a case when a coalition does not breach privacy and otherwise.

**Adaptive Ordering of Adversaries,** In order to facilitate the above pruning in both directions, we adaptively order the coalitions based on their attack powers (Figure 5). This is motivated by the following observations. For downward pruning, super-coalitions of m adversaries with limited attack powers are preferred to check first as they are less likely to breach privacy and hence increase the chance of downward pruning. In contrast, sub-coalitions of m-adversaries with significant attack powers are preferred to check first as they are more likely to breach privacy and hence increase the chance of upward pruning (early-stop).

**The Top-Down Algorithm,** The top-down algorithm checks the coalitions in a top-down fashion using downward pruning, starting from (nG-1) adversaries and moving down until a violation by an m-adversary is detected or all m-adversaries are pruned or checked

**The Bottom-Up Algorithm,** The bottom-up algorithm checks coalitions in a bottom up fashion using upward pruning, starting from 0-adversary and moving up until a violation by any adversary is detected (early-stop) or all m-adversaries are checked.




Downward Pruning Upward Pruning

Fig. 3 Pruning strategies for m-privacy check.

**The Binary Algorithm,** The binary algorithm, inspired by the binary search algorithm, checks coalitions between (nG-1) adversaries and m-adversaries and takes advantage of both upward and downward prunings (Figure 3, Algorithm 1). The goal of each iteration is to search for a pair Isub and Isuper , such that Isub is a direct sub-coalition of Isuper and Isuper breaches privacy while Isub does not. Then Isub and all its sub-coalitions are pruned (downward pruning), Isuper and all its super-coalitions are pruned (upward pruning) as well.

**Adaptive Selection of Algorithms,** Each of the above algorithms focuses on different search strategy, and hence utilizes different pruning. Which algorithm to use is largely dependent on the characteristics of a given group of providers. Intuitively, the privacy fitness score (Equation 1), which quantifies the level of privacy fulfillment of records, may be used to select the most suitable verification algorithm. The higher the fitness score of attacked records, the more likely m-privacy will be satisfied, and hence a top-down algorithm with downward pruning will significantly reduce the number of adversary checks. We utilize such an adaptive strategy in the anonymization algorithm (discussed in the next section) and will experimentally compare and evaluate different algorithms in the experiment section.

**Algorithm 1:** The *binary* $m$-privacy verification algorithm.

```
Data: Anonymized records T* from providers P, an EG
      monotonic C, a fitness scoring function score_F, and the m.
Result: true if T* is m-private w.r.t. C, false otherwise.
sites = sort_sites(P, increasing_order, score_F)
use_adaptive_order_generator(sites, m)
while is_m-privacy_verified(T*, m, C) = false do
    I_super = next_coalition_of_size(n_G - 1)
    if privacy_is_breached_by(I_super, C) = false then
        prune_all_sub-coalitions_downwards(I_super)
        continue
    I_sub = next_sub-coalition_of(I_super, m)
    if privacy_is_breached_by(I_sub, C) = true then
        return false // early stop
    while is_coalition_between(I_sub, I_super) do
        I = next_coalition_between(I_sub, I_super)
        if privacy_is_breached_by(I,C) = true then
            I_super = I
        else
            I_sub = I
    prune_all_sub-coalitions_downwards(I_sub)
    prune_all_super-coalitions_upwards(I_super)
return true
```

## VI. CONCLUSION

In this paper, we considered a new type of potential at-tackers in collaborative data publishing – a coalition of data providers, called m-adversary. To prevent privacy disclosure by any m-adversary we showed that guaranteeing m-privacy is enough. We presented heuristic algorithms exploiting equiv-alence group monotonicity of privacy constraints and adaptive ordering techniques for efficiently checking m-privacy. We introduced also a provider-aware anonymization algorithm with adaptive m-privacy checking strategies to ensure high utility and m-privacy of anonymized data. Our experiments confirmed that our approach achieves better or comparable utility than existing algorithms while ensuring m-privacy efficiently. There are many remaining research questions. Defining a proper privacy fitness score for different privacy constraints is one of them. It also remains a question to address and model the data knowledge of data providers when data are distributed in a vertical or ad-hoc fashion. It would be also interesting to verify if our methods can be adapted to other kinds of data such as set-valued data.

## REFERENCES


[1]. C. Dwork, "Differential privacy: a survey of results," in Proc. of the 5th Intl. Conf. on Theory and Applications of Models of Computation, 2008.

[2]. B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, June 2010.

[3]. C. Dwork, "A firm foundation for private data analysis," Commun. ACM,vol. 54, January 2011.

[4]. N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 4,October 2010.

[5]. W. Jiang and C. Clifton, "Privacy-preserving distributed k-anonymity," in Data and Applications Security XIX, ser. Lecture Notes in Computer Science, 2005.

[6]. W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," VLDB J., vol. 15, 2006.

[7]. O. Goldreich, Foundations of Cryptography: Volume 2, Basic Applications. Cambridge University Press, 2004.

[8]. Y. Lindell and B. Pinkas, "Secure multiparty computation for privacypreserving data mining," The Journal of Privacy and Confidentiality,vol. 1, 2009.

[9]. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in ICDE, 2006.

[10]. P. Samarati, "Protecting respondents' identities in microdata release," IEEE T. Knowl. Data En., vol. 13, 2001.

[11]. L. Sweeney, "k-anonymity: a model for protecting privacy," Int. J.Uncertain. Fuzz., vol. 10, 2002.

[12]. N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and ldiversity," in In Proc. of IEEE 23rd Intl. Conf. on Data Engineering (ICDE), 2007.

[13]. R. Burke, B. Mobasher, R. Zabicki, and R. Bhaumik, "Identifying attack models for secure recommendation," in In Beyond Personalization: A Workshop on the Next Generation of Recommender Systems, 2005.

[14]. D. Kifer, "Attacks on privacy and definetti's theorem," in Proc. of the 35th SIGMOD Intl. Conf. on Management of Data, 2009.

[15]. D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in Proc. of the 2011 Intl. Conf. on Management of Data.

[16]. K. Lefevre, D. J. Dewitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in ICDE, 2006.

[17]. G. Cormode, D. Srivastava, N. Li, and T. Li, "Minimizing minimality and maximizing utility: analyzing method-based attacks on anonymized data," Proc. VLDB Endow., vol. 3, Sept. 2010.

[18]. Y. Tao, X. Xiao, J. Li, and D. Zhang, "On anti-corruption privacy preserving publication," in Proc. of the 2008 IEEE 24th Intl. Conf. on Data Engineering, 2008.

[19]. L. Sweeney, "Datafly: A system for providing anonymity in medical data," in Proc. of the IFIP TC11 WG11.3 Eleventh Intl. Conf. on Database Security XI: Status and Prospects, 1998.


## AUTHORS PROFILE


| | |
|---|---|
| **Author's Name** | : Sanjeev |
| **College Name** | : AMCEC, Bengaluru. |
| **Date of Birth** | : 5th May 1984 |
| **Marital Status** | : Single |
| **Gender** | : Male |
| **Languages Known** | : English, Hindi & Kannada. |