Security of Collaborative Data Publishing in Multiparty Communication

Jayashri K. Bhosle
M.E (CSIT) 2013-2015
M.B.E's COE, Ambajogai
Maharashtra, India

Vanja R. Chirchi
Asst Professor, Dept.
M.B.E's COE, Ambajogai
Maharashtra, India

ABSTRACT
More than one data provider collaborate to publish their data is considered here. m-privacy is a technique proposed to defend m-adversary during collaborative data publishing. M-privacy satisfies the privacy problem while publishing sensitive data. Apart from providing privacy to published data, it is also necessary to provide security between the data provider and third party/un-trusted server, to ensure this, Secure multiparty communication (SMC) protocol is used to provide secure data transfer from publisher and server. There were techniques such as k-anonymity, l-diversity, t-closeness, which were proposed to handle external attacks in data publishing, but none is published for considering internal attacks. This m-privacy is a technique, which considers internal attacks.

General Terms
Binary Algorithm, Heuristic Algorithm, K-Anoymity, t-closness

Keywords
Anoymization, Adversary, TTP, SMC

1. INTRODUCTION
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 [1] ,[2],[3]have 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 [2],[4]. One approach is for each provider to anonymize the data independently, which results in potential loss of integrated data utility. A more desirable approach is collaborative data publishing, [5],[6],[2],[4] which anonymizes data from all providers as if they would come from one source, using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols to do computations[7][8]

Problem Settings . Consider the collaborative data publishing setting 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. Our goal is to publish an anonymized view of the integrated data such that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties. Considering different types of malicious users and information they can use in attacks.

2. EXISTING SYSTEM

2.1 Attacks by External Data Recipient Using Anonymized Data

i. A data recipient, could be an attacker and attempts to infer additional information about the records using the published data and some background knowledge (BK) such as publicly available external data.
ii. Bayes-optimal privacy notion is used to protect against specific types of attacks by assuming limited background knowledge.
iii. For example, k –anonymity [10],[11] prevents identity disclosure attacks by requiring each equivalence group, records with the same quasi-identifier values, to contain at least k records.
iv. Representative constraints that prevent attribute disclosure attacks include l-diversity, which requires each equivalence group to contain at least l "well-represented" sensitive values[9].
v. t-closeness[12], which requires the distribution of a sensitive attribute in any equivalence group to be close to its distribution in the whole population.
vi. Differential privacy[1][3] publishes statistical data or computational results of data and gives unconditional privacy guarantees independent of attackers background knowledge.

2.2 Attacks by Data Providers Using Intermediate Results and Their Own Data

i. The data providers are semihonest, commonly used in distributed computation setting. They can attempt to infer additional information about data coming from other providers by analyzing the data received during the anonymization.
ii. 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.

Disadvantages

i. TTP or SMC do not protect against data providers to infer additional information about other records using the anonymized data and their own data

3. PROPOSED SYSTEM

3.1 Attacks by Data Providers Using Anonymized Data and Their Own Data

i. Collaborative data publishing setting with horizontally partitioned data across multiple data providers, each contributing a subset of records is considered.
ii. A data provider could be the data owner itself who is contributing its own records.
iii. 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.

Advantages

i. "Insider attack" by data providers is considered.

4. ANALYSIS RESULT

i. m- privacy Verification

M	Power of m-privacy
N	No of data providers
Ng	Number of data providers contributing to a group
Tg	Number of records in a group

Table: Experimental Parameter

Compared the different m-privacy verification heuristics against different attack powers, used two different groups of records with relatively small and large average number of records per data provider, respectively. Figure 1 and 2 shows the runtime with varying m for different heuristics for the two groups.

Table I.

M (Power of m-privacy)	Binary	Top-Down	Direct	Bottom Up
5	2	20	4	5
6	4	15	4	10
7	8	10	4	15
8	12	5	5	20

Tg/Ng=10

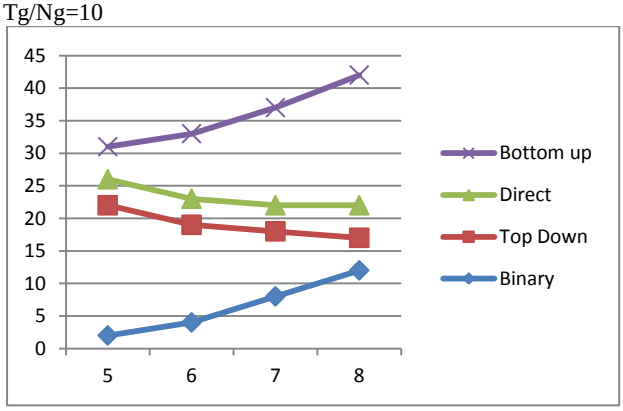


Fig.1. Runtime (logarithmic Scale) vs m(power of m-privacy)

Table II.

M (Power of m-privacy)	Binary	Top Down	Direct	Bottom Up
15	500	500	300	100
20	400	300	300	150
25	200	150	300	300
30	300	100	300	500

Tg/Ng=50

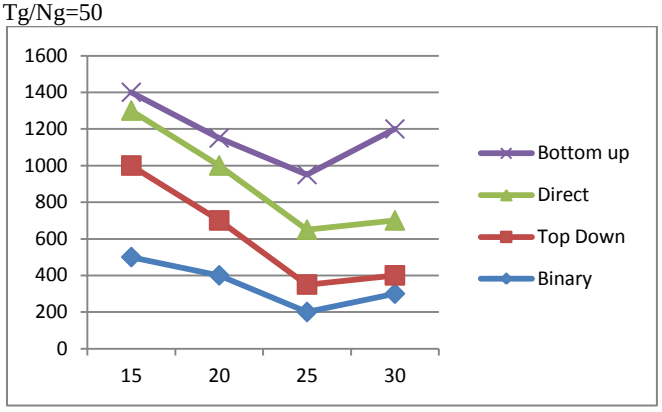


Fig. 2. Runtime (logarithmic Scale) Vs m

Analyzed the impact of contributing data providers (nG)on the different algorithms for the small and large group respectively. Figure 3 and 4 shows the runtime of different heuristics with varying number of contributing data providers nG.

Table III.

N(No. of Data Providers)	Binary	Top Down	Direct	Bottom Up
3	2	14	4	4
6	4	12	5	8
9	6	8	6	12
12	8	4	7	14

Tg/Ng=10

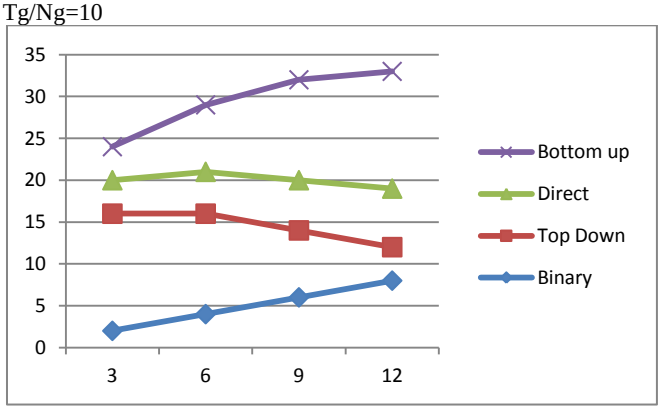


Fig.3. Runtime (logarithmic Scale) Vs no. of Data Providers

Table IV.

N (No. of Data Providers)	Binary	Top Down	Direct	Bottom Up
150	40	500	100	200
300	40	400	100	300
450	40	300	100	400
600	60	200	200	500

Tg/Ng=50

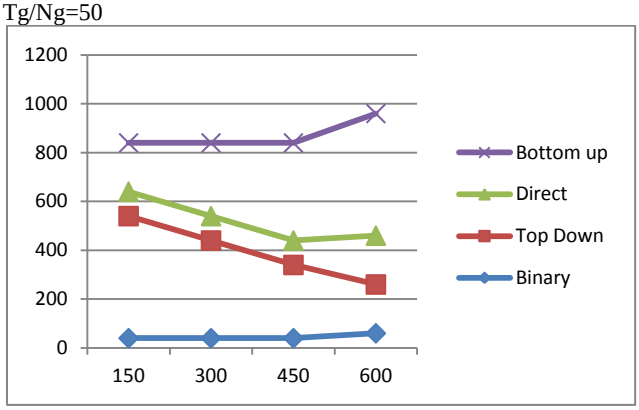
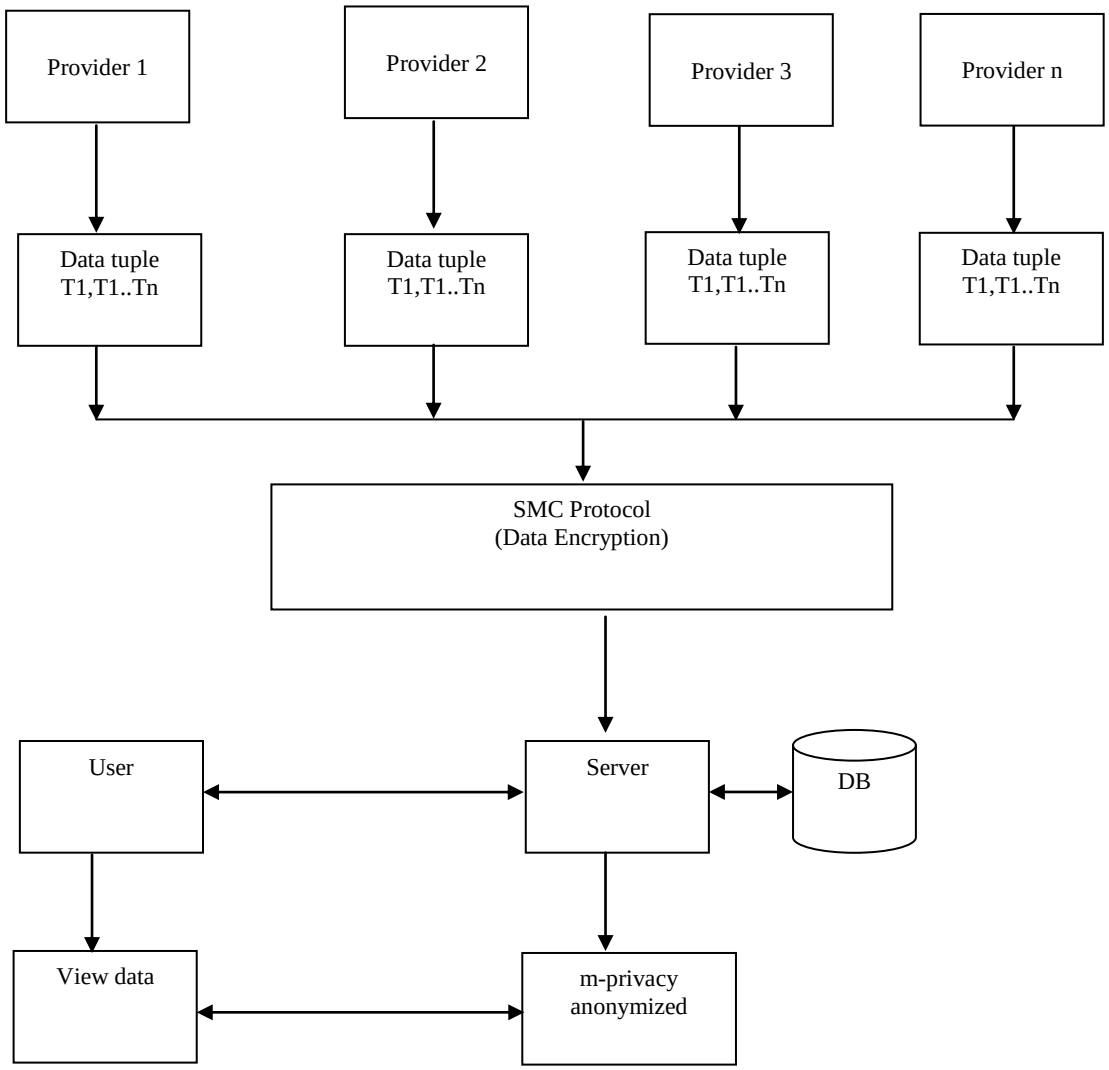


Fig.4. Runtime (logarithmic Scale)Vs n (no of data providers)

5. SYSTEM ARCHITECTURE



6. FUTURE SCOPE

A new type of potential attackers in collaborative data publishing – a coalition of data providers, called m-adversary is considered. To prevent privacy disclosure by any m-adversary we showed that guaranteeing m-privacy is enough. A heuristic algorithm is presented exploiting equivalence 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.

7. 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, pp. 1–19.

[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, pp. 14:1–14:53, June 2010.

[3] C. Dwork, "A firm foundation for private data analysis," Commun. ACM, vol. 54, pp. 86–95, 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, no. 4, pp. 18:1–18:33, 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, vol. 3654, pp. 924–924.

[6] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," VLDB J., vol. 15, no. 4, pp. 316–333, 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 privacy preserving data mining," The Journal of Privacy and Confidentiality, vol. 1, no. 1, pp. 59–98, 2009.

[9] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,-diversity: Privacy beyond k-anonymity," in ICDE, 2006, p. 24.

[10] P. Samarati, "Protecting respondents' identities in microdata release," IEEE T. Knowl. Data En., vol. 13, no. 6, pp. 1010–1027, 2001.

[11] L. Sweeney, "k-anonymity: a model for protecting privacy," Int. J.