Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Distributed Rule Mining Under Privacy Ensured Multiparty Environment

D.Sudhapriya
Research Scholar Department of Computer Science
SSM College of Arts & Science
Komarapalayam, India

Mr.A.Sridhar
Head & Professor
SSM College of Arts & Science
Komarapalayam, India

Abstract-Secure multiparty protocols have been proposed to enable non-colluding parties to cooperate without a trusted server. Even though such protocols prevent information disclosure other than the objective function, they are quite costly in computation and communication. The high overhead motivates parties to estimate the utility that can be achieved as a result of the protocol beforehand. In this paper, we propose a look-ahead approach, specifically for secure multiparty protocols to achieve distributed k-anonymity, which helps parties to decide if the utility benefit from the protocol is within an acceptable range before initiating the protocol. The look-ahead operation is highly localized and its accuracy depends on the amount of information the parties are willing to share. Experimental results show the effectiveness of the proposed methods.

Keywords- Secure multiparty computation, distributed k-anonymity, privacy, security.

I. Introduction

Secure multiparty computation (SMC) protocols are one of the first techniques used in privacy preserving data mining in distributed environments. The idea behind these protocols is based on the theoretical proof that two or more parties, both having their own private data, can collaborate to calculate any function on the union of their data. While doing so, the protocol does not reveal anything other than the output of the function or anything that can be computed from it in polynomial time. Moreover, the protocol does not require a trusted third party. While these properties are promising for privacy preserving applications, SMC may be prohibitively expensive. In fact, many SMC protocols for privacy preserving data mining suffer from high computation and communication costs. Furthermore, those that are closest to be practical are designed for the semi honest model, which assumes that parties will not deviate from the protocol. Theoretically, it is possible to convert protocols in the semi honest model into protocols in the malicious model. However, the resulting protocols are even more costly.

II. Secure Look Ahead

The parties have private inputs $T_{1,}$ . . . , T_n and wish to securely compute a function *f* on these inputs. Briefly, a protocol $\prod_f$ for computing *f* is a set of Turing machines (one per party). The Turing machines are connected pair wise with communication tapes on which they can send and receive (private) messages. A protocol is executed by running the Turing machines where each Turing machine gets the private input of the corresponding party, we write $\prod_f$ [T_1, . . . , T_n]_. The list of all messages received by the *i*th Turing machine on all its communication tapes during the execution of the protocol is called the *view* of party *i*. A protocol for functionality *f* is said to be secure if, for all parties *i*, the view of party *i* can be efficiently *simulated* from the input, T_i the output *f*($T_{1,}$ . . . , T_n)[1] and any background information of the *i*th party. Put simply, simulating the view of party *i* means that there exists a polynomial time probabilistic algorithm which can generate tuples with a statistical distribution that is indistinguishable (either statistically close, or computationally indistinguishable in polynomial time, depending on the desired security model) from the view of the party.

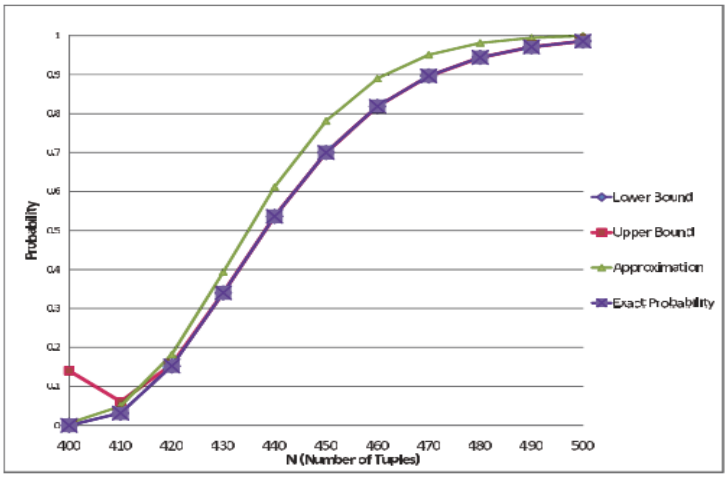
III. Problem Definition

Distributed k-anonymity protocol is *c,p*-sufficient for party σ iff Note that we take the safe road and require one mapping in the set S_ to have a sufficient probability of producing k anonymization. This is not a necessary requirement as the probability of producing a k-anonymization from some mapping in S_ is always higher than producing it from any given mapping within the set.

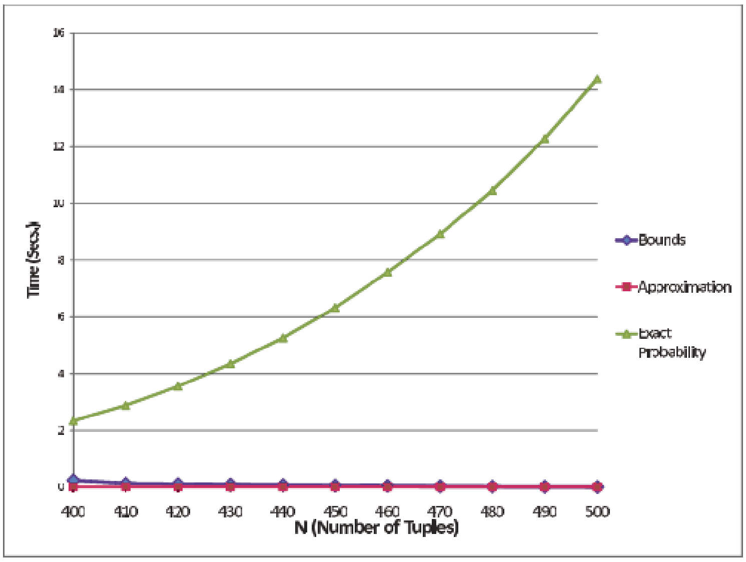
IV. Experiments and Analysis

The experiments aim to convey the accuracy of the probability approximation and the time performance with respect to different variables such as data size and the value of k.

A) Synthetic data Set: For the synthetic data set, we set k to 100 and the number of equivalence classes to 4, which means each equivalence class, will contain 25 percent of the total number of tuples due to the uniform distribution assumption.

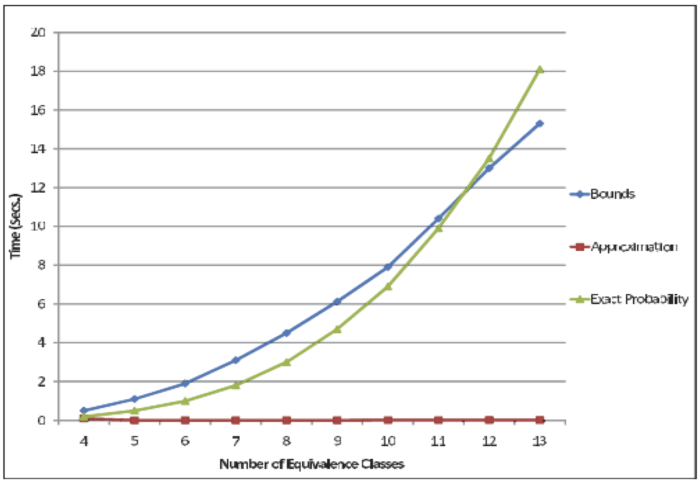


This Figure shows that Probability results on synthetic data set.



This graph shows Time performance on synthetic data set.

B) Shuffled Adult Data Set: To create the data sets for each site, we have partitioned the data set as follows: First, we have selected and removed 15 percent of the data (which corresponds to 4,524 tuples) to form the data of the home site. Then, using the remaining 85 percent of the data, by random sampling, we have formed the data of the remote site which has a size equal to the home site. Although home site knows the general distribution of the data of the remote site, data itself are invisible to the home site. Repeating the latter step for 50 times we had 50 different remote sites that are going to be subject to Look Ahead based on the data of the home site. Conducting several experiments on randomized data gives us an idea on the algorithm behavior at the mean.



The Graph Shows Time performance on synthetic data set.

The below figure shows the accuracy of the Randomized Approximation for each data size mentioned above and for different k values. We use the adjective "Randomized" when the remote site shares a distorted version of the atomic frequencies instead of the actual distribution, to bound the information released by its local k-anonymization.

C) DPK Protocols: In this set of experiments, we use the original Adult data set, which has attribute correlations. The data sets are obtained in the same manner with the ones in Section 8.2 but this time we did not shuffle them. Also we use bigger data sizes for the home and the remote sites which are 30, 40, and 50 percent.

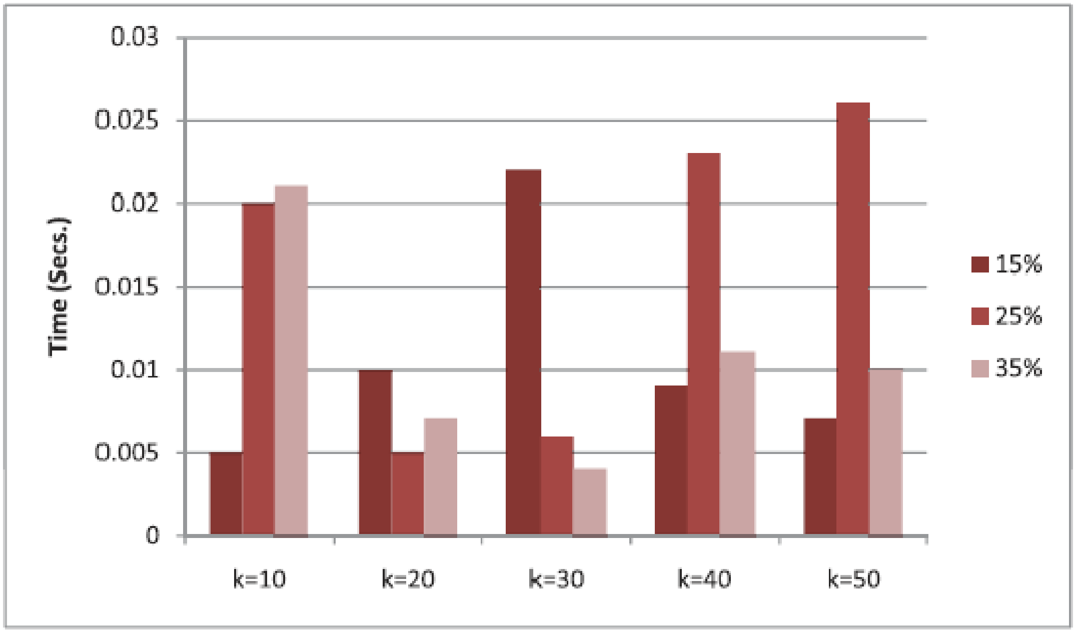


Fig: Time performance on shuffled data set.

The Below Figure shows the error percentage with respect to data size, value of k and c when we fix threshold p at 0.9. In most cases, the error rate is no bigger than 0.15. Especially, for mappings further away from the mapping of the local anonymization.

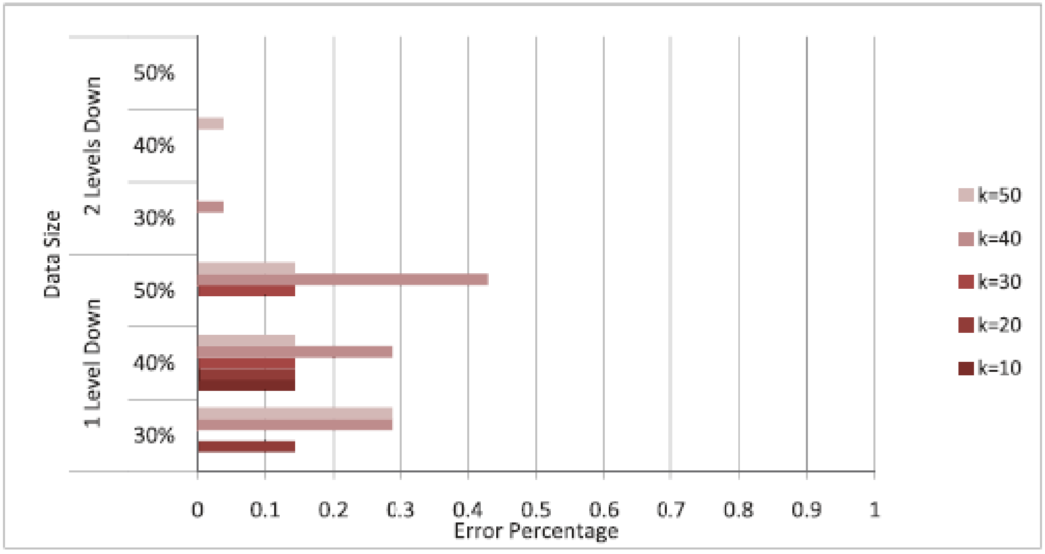


Fig: Error percentage on adult data set for different k values.

V. Conclusion And Future Work

Most SMC protocols are expensive in both communication and computation. We introduced a look-ahead approach for SMC protocols that helps involved parties to decide whether the protocol will meet the expectations before initiating it. We presented a look-ahead protocol specifically for the distributed k-anonymity by approximating the probability that the output of the SMC will be more utilized than their local anonymization. Experiments on real data showed the effectiveness of the approach. Designing look ahead for other SMC protocols stands as a future work. A wide variety of SMC protocols have been proposed especially for privacy preserving data mining applications each requiring a unique look ahead approach. As for the look-ahead process on distributed anonymization protocols, definitions of k-anonymity definitions can be revisited, more efficient techniques can be developed and experimentally evaluated.

References:

1) N. Li and T. Li, "T-Closeness: Privacy Beyond K-Anonymity and L-Diversity," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE '07), Apr.2007.
2) D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, "Worst-Case Background Knowledge for Privacy-Preserving Data Publishing," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE '07), Apr. 2007.
3) A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "'-Diversity: Privacy beyond k-Anonymity," Proc. IEEE 22nd Int'l Conf. Data Eng. (ICDE '06), Apr. 2006.
4) M. Kantarclu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data,"IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037, Sept. 2004.
5) P. Samarati, "Protecting Respondent's Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.