Privacy Preserving Approaches for Global Cycle Detections for Cyclic Association Rules in Distributed Databases

Nirali R. Nanavati and Devesh C. Jinwala
Computer Engineering Department, Sardar Vallabhbhai National Institute of Technology, Surat, India

Keywords: Privacy, Cyclic Association Rules, Distributed Setup.

Abstract: The current massive proliferation of data has led to collaborative data mining that requires preservation of individual privacy of the participants. A number of algorithms proposed till date in this scenario are limited to mining association rules and do not consider their cyclic nature that finds associations with respect to the time segment. Hence catering to this challenge, we propose techniques for privacy preservation while finding global cycles when mining cyclic association rules in a distributed setup. The proposed techniques are based on homomorphic encryption and Shamir's secret sharing that can help us decipher partial and total global cycles along with maintaining privacy in a distributed setup. Additionally security, efficiency and correctness analysis of the proposed algorithms are also given.

1 INTRODUCTION

The concept of coopetation (Pedersen and Saygn, 2007) which symbolizes a blend of competition and cooperation has become an important practice for various organizations to sustain and succeed in this competitive world. The approach of coopetation has its application in Distributed Data Mining in which privacy needs of the participating parties are to be guaranteed. That is, the data mining results of each of the sites that satisfy a certain function only, must be known in the cumulative data. The identities of the participants and the remaining data that does not satisfy the cumulative function must be kept secret.

The solution for efficient implementation of privacy preserving techniques for distributed data mining is one of the most researched and challenging fields of study today.(Venkatadri and Reddy, 2011)

A major limitation of the existing distributed privacy preserving algorithms in this area (Kargupta et al., 2007); (Kantarcioglu and Clifton, 2004); (Kantarcioglu, 2008); (Vaidya, 2008); (Ge et al., 2010); (Aggarwal and Yu, 2008); (Sang et al., 2009); (Shi et al., 2011) is that none of them deal with cyclic data or data that is segmented and temporal.

The paper (Sang et al., 2009) does mention privacy preserving tuple attribute matching in a distributed setup; but again, it does not deal with cyclic association rules. In addition, the recent work by Shi et al. (Shi et al., 2011) does explain approaches for aggregation of time series based data using a third party aggregator with its main application in wireless sensor networks; but then it does not cater to cyclic association rules.

In a number of real life applications, data is time based. So it is important to decipher the rules that are frequent and repeat at regular time intervals.In this context, we propose a construction comprising of techniques to find global cycles among cyclic association rules (Ozden et al., 1998); (Ben and Gouider, 2010) (which is a subset of temporal rules) while maintaining the privacy in a co-opetative model. Our construction is restricted to a horizontally partitioned homogeneous model. Any scenario, where there are coopetative parties involved and the data to be mined is segmented on the basis of time, can use the approaches we propose to exchange their information and preserve their privacy as well.

The two proposed techniques for the multi-party scenario are based on Efficient Private Matching (EPM) (Freedman et al., 2004) and Shamir's secret sharing technique (Shamir and Adi, 1979) respectively to find global cycles privately. These methods have their pros and cons which have been discussed in this paper.

R. Nanavati N. and C. Jinwala D..
Privacy Preserving Approaches for Global Cycle Detections for Cyclic Association Rules in Distributed Databases.
DOI: 10.5220/0004018803680371
In *Proceedings of the International Conference on Security and Cryptography* (SECRYPT-2012), pages 368-371
ISBN: 978-989-8565-24-2

An important example of such a scenario where privacy preservation is required while mining global cycles would be a group of television channels who plan to share their data like their Television Rating Points (TRPs), timings of the various programmes being watched and various advertisements being shown, by maintaining their privacy as well, so as to figure out which TV program is being watched at what time periodically by relatively more number of viewers. The important thing to note here is that the competing channels do not want to disclose their identity and still be able to find out global cycles (i.e. trends that repeat at regular intervals of time) amongst every channel's data.

2 THE PROPOSED ALGORITHM

We propose an algorithm that finds cyclic association rules (Ozden et al., 1998) using the interleaved algorithm that are supported at all the participating parties privately. The problem considers a co-opetative scenario of homogenous databases where there are 'p' semi honest parties collaborating to find the global cycles (cyclic and frequent association rules in a particular time segment at all the participating parties). Let there be a set of 't_i' transactions and maximum N items at each of the partitions (where $0 < i \leq p$) and each transaction has a subset of 'N' items.

Let the schema at each of the sites be of the form <tid, items, timestamp>. All the parties decide on the local minimum Support 'minSupp', minimum Confidence 'minConf' and the length of cycle which is less than or equal to the set of transactions at each of the parties. They also decide on one of the privacy preserving protocols (explained below) that they would follow.

In the first protocol i.e. Efficient Private Matching algorithm (Freedman, Nissim and Pinkas, 2004), for the multiparty scenario, we consider p parties, $P_1 \ldots P_p$, with corresponding lists of inputs $X_1 \ldots X_p$ which are the cyclic rules at each party. We assume each list contains k_i ($0<i\leq p$) inputs. The parties would be able to compute the intersection of all p lists privately using homomorphic encryption as explained in (Freedman, Nissim and Pinkas, 2004).

The second protocol is based on Shamir's secret sharing (Shamir and Adi, 1979); (Ge et al., 2010). The algorithm finds the sum of the secret information (which will help us decipher the global cycles in our scenario privately) in the cumulative data.

Algorithm 1: Finds the sum of the cyclic rules at all the parties privately using Shamir's secret sharing
Input: *N, k, p,* S_{ijt}(secret values)
1: **for** each transaction j = 1 to N **do**
2. **for** each time segment t = 1 to k **do**
3: **for** each party P_i, (i = 1, 2 ,.., p) **do**
4: each party selects a random polynomial $q_i(x) = A_{p-1}x^{p-1} + \ldots + a_1x^1 + S_{ijt}$
5: compute the share of each party P_t, where share(S_{ijt}, P_t) = $q_i(x_t)$
6: **for** t = 1 to p **do**
7: send share(S_{ijt}, P_t) to party P_t
8: receive the shares share(S_{ijt}, P_t) from every party P_t.
9: **end for**
10: compute $Sum(x_i) = q_1(x_i) + q_2(x_i) + .. ,+q_p(x_i)$
11: **for** t = 1 to p **do**
12: send $Sum(x_i)$ to party P_t
13: receive the results $Sum(x_i)$ from every party P_t
14: solve the set of equations to find the sum of $\sum_{i=1}^{p} S_{ijt}$ secret values.
15: **end for**
16: **if** ($\sum_{i=1}^{p} S_{ijt}$ = p) globalCycle = 1;
17: **else if** ($\sum_{i=1}^{p} S_{ijt} < p$) ; cnt (partialCycle) = $\sum_{i=1}^{p} S_{ijt}$
18: **end for end for end for**

Figure 1: Secret sharing approach applicable to our scenario.

There are 2 possible variations that can also be applied to our approach:

1. If the co–opetative setup decides on finding global association rules instead of global cycles; if the association rule is cyclic; only then the count of the item is exchanged privately (we could use Secret Sharing or Secure Sum algorithm (Kantarcioglu, 2008) for this) by which we can calculate the global support and hence find global association rules.

2. If partial cycles (cyclic association rules supported at some of the parties) above a certain threshold need to be determined; then the Secret sharing approach could be used to determine privately the count of the parties that support a certain cyclic rule.

3 ANALYSIS

We begin by doing the correctness analysis of our proposed protocols followed by the complexity and

security analysis. To find the globally cyclic association rules; assume that the party P_i has a private vector V_i for a particular cyclic association rule C_j (j is the total number of rules possible). This vector V_i has bits for whether a particular rule is cyclic or not in a time segment 'k' such that $1<k\leq t$(t is the total no. of time segments). Hence for a particular rule C_j; we try and find the $\sum_{i=1}^{p} V_{ik}$ for a particular cyclic association rule in a particular time segment across different parties.

Algorithm 2: *To decipher global cycles in cyclic association rules among different parties*
Input: *p, l, minSupp, minConf,* t_i, N_i
1: **for** i = 1 to p **do**
2: itemSet$_i$:= ϕ ; *//set of all frequent itemsets*
3: sub$_i$ = genSubsets (N_i);
4: **for** items it = 1 to 2^N-1 **do**
//2^N-1 is no. of subsets.
5: **if** (cardinality(sub$_{it}$)>1 && all elements of sub$_{it}$ ∉ itemSet$_i$)
//cycle skipping and elimination is applied
6: **break;**
7: **end if**
8: **for** j =1 to t_i **do**
9: **for** time segment m = 1 to k
10: **if** cyclic(sub$_{it}$) for time segment m
11: **if** (support(sub$_{it}$)>minSupp)
12: itemSet$_i$ U {sub$_{it}$};
13: count ++;
14: **end if end if end for end for**
15: frequentRule$_i$:= ϕ ;
16: **for** each element x in itemSet$_i$ **do**
17: rule := newRule$_i$(x);
18: **if** conf(rule) > minConf)
19: frequentRule$_i$: = frequentRule$_i$ U rule
20: **end if end for end for**
//Association rules are formed at each party
21: globalCycle := ϕ ;
22: **for** m = 1 to k (k = no. of time segments)
*/*Each setup will decide on the privacy preserving protocol they will use: Efficient Private Matching or Secret Sharing*/*
23: **if** IsGloballyCyclic(frequentRule)
//for all parties i; this info is determined privately
24: globalCycle:=globalCycleU frequentRule;
25: **end if end for End;**

Figure 2: Algorithm to find global cycles privately in a distributed setup.

Let us consider the protocol 1 based on efficient private matching. Since this scenario has (p-1) partitions and one leader S; and for ∀ y ∈ S;the leader prepares p-1 shares that XOR to y. Now each party formulates a polynomial equation whose roots are their respective inputs. The coefficients of these polynomials are encrypted and sent to the leader.

The leader takes Enc($r_y \cdot P(y_i') + y_i'$) and sends the users the Enc(y_i') where y_i' is one of the shares that XOR to y (y ∈ S). Hence all the p – 1 parties will learn their share of Enc(y_i'). The parties decrypt their share and XOR it with the other parties. They will learn y only if $y \in P_1 \cap P_2 \cap P_3 \cap \ldots$.S. Hence we will be able to find the cyclic association rule that is supported by all the parties privately.

On the other hand, considering the protocol 2 using the secret sharing, we need to solve the linear equations formulated which have p unknown coefficients and n equations. Since the polynomials would have a unique solution; each party P_i can solve the equation and find the value of $\sum_{i=1}^{p} V_{ik}$ for a particular cyclic association rule. However it cannot determine the secret values of the other parties as the polynomial coefficients of each party are hidden from the other.

If the $\sum_{i=1}^{p} V_{ik} = p$; it implies that a global cycle does exist among all parties for C_j in the time segment k. If on the other hand $\sum_{i=1}^{p} V_{ik} = z$ (z < p) ; it implies that a partial cycle exists at z parties.

Similarly other global cycles can be deciphered privately with respect to time with the condition that all parties follow the protocol honestly.

For the complexity analysis, consider first the complexity of the cyclic association rules generation. If we have n items, x time segments and t_i transactions ($0 < i \leq p$) such that p is the no. of partitions; then the number of possible association rules with respect to time is $O(x*n*2^{n-1})$. So the computation complexity of the cyclic association rule generation would be $O(n*\max(t_i)*2^n*x)$.

For the efficient private matching algorithm; the communication overhead is O (pk) where there are p-1 parties and one leader and k coefficients sent using homomorphic encryption to the leader. As far as the computation cost is concerned it involves encryption and decryption of k values at each partition site and for the leader to compute p-1 shares that XOR to y and Enc ($r_y \cdot P(y_i') + y_i'$) using k exponentiations and hence total $O(k^2)$. Thus, this protocol has a much lesser communication cost compared to the Protocol 2 that uses Shamir's secret sharing technique.

Secondly, for the secret sharing algorithm with 'p' parties; each party sends its shares to 'p-1' other parties and if we consider 'k' such bits based on the no. of transactions and the no. of time segments; the complexity would be $O(kp^2)$.

Along with the communication cost there would also be the computation cost of generating the random polynomial, p^2 computations, p(p-1) additions, and solving the equations with unknowns

to find the sum $\sum_{i=1}^{p} V_{ik}$.

Regarding the security analysis, our protocol that finds global cycles based on efficient private matching is semantically secure from attackers and also preserves privacy. The data of all the clients is private as the leader only sees the encrypted coefficients. The leader's privacy is also protected as he only sends across encrypted shares that XOR to y and hence the actual value of y is not known to the different participants.

Our protocol based on secret sharing, is also semantically secure from the network attackers. i.e. network attackers cannot learn any valuable information except for the shares of each of the parties which would be of no use. This algorithm can also effectively prevent collusive behaviour if the number of collusive parties c < p-1.

4 CONCLUSIONS

In this paper we define a new problem of detecting partial and total global cycles in a co-opetative setup while maintaining the privacy of the individual parties.

We extend the interleaved algorithm to find global cycles in cyclic association rules privately. Our first algorithm has higher computation cost and lesser communication cost compared to the second one based on Shamir's Secret sharing.

However there are a few open research challenges that which include applying these privacy preserving theories to other temporal rule mining methods like calendric association rules, temporal predicate association rules, OLAP cubes and sequential association rules. The algorithm could also be extended to a heterogeneous setup and to malicious models. Since our approach considers the threshold count at each of the parties individually for a particular item and gives equal importance to all the participants; an extension to this could be to privately decipher global cyclic association rules considering the global count of the respective item and hence giving importance to the parties with respect to their transaction data size.

REFERENCES

Aggarwal, C. C. and Yu, P. S., (2008). A general survey of privacy-preserving data mining models and algorithms. In *Privacy-Preserving Data Mining*, volume 34 of The Kluwer International Series on Advances in Database Systems, pages 11–52. Springer US.

Ben Ahmed, E. and Gouider, M. S., (2010). Towards a new mechanism of extracting cyclic association rules based on partition aspect. In *Research Challenges in Information Science (RCIS), 2010 Fourth Interna tional Conference on*, pages 69 –78.

Freedman, M. J., Nissim, K., and Pinkas, B., (2004). Efficient private matching and set intersection. Pages 1–19. *Springer-Verlag*.

Ge, X., Yan, L., Zhu, J., and Shi, W., (2010). Privacy preserving distributed association rule mining based on the secret sharing technique. In *Software Engineer ing and Data Mining (SEDM), 2010 2nd International Conference on*, pages 345 –350.

Kantarcioglu, M., (2008). A Survey of Privacy-Preserving Methods Across Horizontally Partitioned Data. In *Privacy-Preserving Data Mining*, volume 34 of Ad-vances in Database Systems, pages 313–335. Springer US.

Kantarcioglu, M. and Clifton, C., (2004). Privacy preserving distributed mining of association rules on horizontally partitioned data. *Knowledge and Data Engineering, IEEE Transactions on*, 16(9):1026 1037.

Kargupta, H., Das, K., and Liu, K., (2007). Multi-party, privacy-preserving distributed data mining using a game theoretic framework. In *Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2007*, pages 523–531, Berlin, Heidelberg. Springer-Verlag.

Ozden, B., Ramaswamy, S., and Silberschatz, A. (1998). Cyclic association rules. In *Proceedings of the Four-teenth International Conference on Data Engineering, ICDE '98*, pages 412–421, Washington, DC, USA. IEEE Computer Society.

Pedersen, T. B., Saygin, Y., and Savas, E. (2007). Secret Sharing vs. Encryption-based Techniques For Privacy Preserving Data Mining. *Sciences-New York*, (December):17–19.

Sang, Y., Shen, H., and Tian, H., (2009). Privacy preserving tuple matching in distributed databases. *IEEE Trans. on Knowl. and Data Eng.*, 21:1767–1782.

Shamir, A., (1979). How to share a secret. *Commun. ACM*, 22:612–613.

Shi, E., Chan, T.-H. H., Rieffel, E. G., Chow, R., and Song, D., (2011). Privacy-preserving aggregation of timeseries data. In *NDSS*.

Vaidya, J. (2008). A survey of privacy-preserving methods across vertically partitioned data.

Venkatadri.Mand Reddy, D. L. C., (2011). Article: A review on data mining from past to the future. *International Journal of Computer Applications*, 15(7):19–22.