A BINARY ESTIMATION AND CONDITIONAL EXPONENTIAL METHODS TO IDENTIFY FREQUENT SUBGRAPHS

$_1$N.Asha Jyothi, $_2$D. Srinivas

[1,2]Dept. of CSE, Kakinada Institute of Engineering & Technology., Matlapalem, Talarevu Mandal, Corangi, E.G.dt, AP, India

ABSTRACT:

We study the issue of frequent subgraph mining (FSM) under the thorough differential security model. We present a two-stage differentially private FSM algorithm, which is alluded to as DFG(differentially private frequent subgraph). We first build a lattice for the recognized frequent subgraphs dependent on their inclusion relations, where every node represents an identified frequent subgraph. To register the noisy sup-port of the graphs represented to by the nodes in a given path in the lattice, we at that point devise a check collection technique, where the noisy help of the graphs is acquired by gathering the counts of the graphs in commonly disjoint subdatabases. We show that, compared with directly perturbing the support of the graphs, our count accumulation strategy can fundamentally improve the precision of the loud backings.

KEYWORDS: FSM, DFG, graphs, noisy

1] INTRODUCTION:

The work proposed by Shen et al. [5]. They utilize the Markov Chain Monte Carlo (MCMC) sampling to extend the exponential mechanism [6], and use the extended exponential mechanism to directly select frequent subgraphs from all possible graphs which may or may not exist in the input graphs. Since confirming the convergence of MCMC stays an open issue when the distribution of tests isn't recognizable, the algorithm proposed in [5] ensures just the more vulnerable (ϵ, δ)-differential protection. As the output space contains every conceivable graphs, it brings about an enormous output space, which makes the selections of frequent subgraphs inaccurate. A developing number of studies have as of late been proposed for mining continuous itemsets and frequent sequences under differential security. Since these investigations are planned explicitly for the sort of pattern being mined, they can't be applied to mining frequent subgraphs. To our best information, we don't know about any current examinations which can discover frequent subgraphs with high data utility while fulfilling ϵ-differential protection.

2] LITERATURE SURVEY:

2.1] M. Kuramochi a ***et al***

Throughout the long term, frequent itemset revelation algorithms have been utilized to discover fascinating patterns with regards to different application zones. As data mining procedures are by and large progressively applied to nontraditional domains, existing frequent pattern revelation approaches can't be utilized. This is on the grounds that the transaction framework that is accepted by these algorithms can't be utilized to successfully model the data sets in these domains. A substitute method of modeling the objects in these informational indexes is to represent to them utilizing graphs. Inside that model, one method of figuring the frequent pattern revelation issue is that of finding subgraphs that occur frequently over the whole set of graphs. We present a computationally proficient algorithm, called FSG, for discovering all frequent subgraphs in enormous graph data sets.

2.2] D. Proserpio ***et al***

We present a way to deal with differentially private calculation in which one doesn't scale up the extent of noise for testing queries+, yet rather downsizes the commitments of testing records. While downsizing all records consistently is equal to scaling up the commotion size, we show that scaling records non-consistently can bring about significantly higher accuracy by bypassing the most pessimistic scenario necessities of differential security for the noise magnitudes.

3] PROBLEM DEFINTION:

FSM means to find all subgraphs that happen in input graphs more much of the time than a given limit.

FSM has handy importance in various applications, running from bioinformatics to social network analysis.

For instance, finding successive subgraphs in informal communities can be crucial to comprehend the mechanics of social interactions.

4] PROPOSED APPROACH:

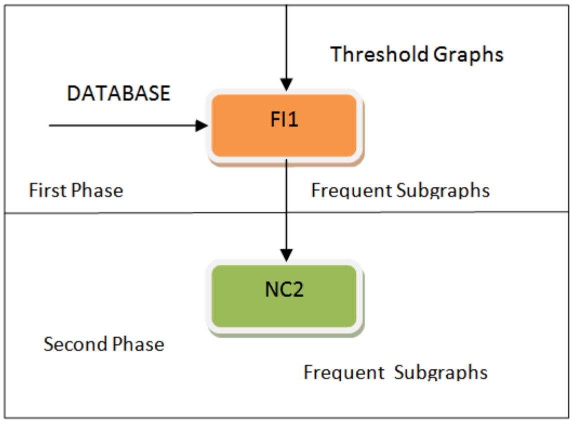
We propose a basic and effective strategy, specifically error-aware path construction, for building the set of paths which are taken as input to the count aggregation technique.

Contrasted with our past work, which needs to utilize two strategies (i.e., the path construction and path extension methods) to get the last set of paths, we utilize just this new strategy to build the set of paths.

Since the error-aware path construction technique can legitimately lreduce the errors of noisy supports during calculation, it brings about preferable information utility over the two strategies proposed in our past work, particularly when the edge is generally low.

Second, to show the generality of our DFG algorithm, we extend it for mining both frequent itemsets and frequent sequences.

5] SYSTEM ARCHITECTURE:



6] PROPOSED METHODOLOGY:

6.1] Admin

The admin can login directly with the application and the admin can perform out the activities like view and approve the user, transfer dataset, view dataset, Manage Item sets, Frequent Item sets.

6.2] User

The user should enroll with the application and the user can login with the application and the user after his/her login perform out certain activities like view profile, perform some actions like view profile, pruning items, view graph.

6.3] Dataset

Dataset contains collection of data. In the case of tabular data, a dataset corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the dataset in question.

6.4] Frequent Itemset

A frequent itemset contains an itemset whose support is greater than some user-specified minimum support.

6.5]FSM

FSM find all subgraphs that occur in input graphs more frequently than a given threshold.

7] ALGORITHM:

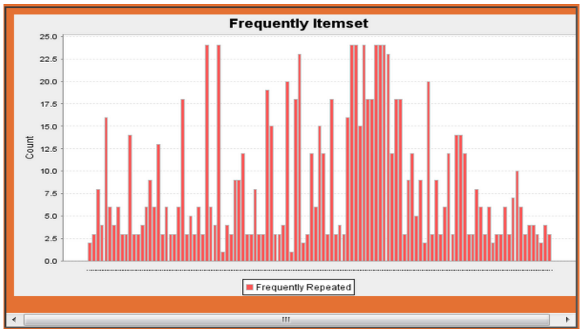
DFG Algorithm

Step 1: In the first phase of DFG, we put forward a frequent subgraph identification approach (referred to as FI1) to privately identify frequent sub graphs in order of increasing size.

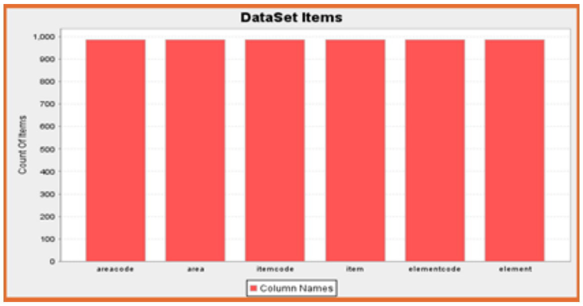
Step 2: A binary estimation method is used to estimate the number of frequent subgraphs.

Step 3: In the second phase of DFG, we devise a lattice-based noisy support computation approach (referred to as NC2) to compute the noisy support of identified frequent subgraphs.

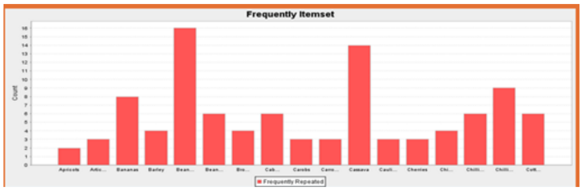
8] RESULTS:



This graph shows that number of frequent itemset present in the given dataset, frequently repeated items. x-axis is frequently repeated items and y-axis is count of frequently repeated items.



Graph shows that raw data(main graph) present in the dataset. It shows the all data present in the data set.



Graph shows that sub graph, This is derived from main graph.

9] CONCLUSION:

The problem of designing a FSM algorithm, which can satisfy ϵdifferential privacy and achieve high data utility. We present a differentially private FSM algorithm called DFG, which consists of a frequent subgraph identification phase and a noisy support computation phase. DFG can be easily extended for mining other frequent patterns, such as frequent itemsets and frequent sequences. Through privacy analysis, we prove that our DFG algorithm satisfies ϵ-differential privacy.

10] EXTENSION WORK:

We extend our DFG algorithm to mine frequent sequences under ϵ-(sum of) differential privacy. We indicate this extended algorithm by DFS. We analyze DFS against a best in class algorithm called PFS, which secretly finds the sequences whose help surpasses a given edge by means of sampling-based applicant pruning. In the analyses, we additionally utilize two real dataset.

11] REFERENCES:

[1] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering frequent patterns in sensitive data," in KDD, 2010.

[2] C. Dwork, "Differential privacy," in ICALP, 2006.

[3] L. Sweeney, "k-anonymity: A model for protecting privacy," Int. J. Uncertain. Fuzziness Knowl.-Base Syst, 2002.

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

[5] E. Shen and T. Yu, "Mining frequent graph patterns with differential privacy," in KDD, 2013.

[6] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in FOCS, 2007.

[7] N. Li, W. Qardaji, D. Su, and J. Cao, "Privbasis: frequent itemset mining with differential privacy," in VLDB, 2012, pp. 305–316.

[8] C. Zeng, J. F. Naughton, and J.-Y. Cai, "On differentially private frequent itemset mining," in VLDB, 2012.

[9] S. Xu, S. Su, X. Cheng, Z. Li, and L. Xiong, "Differentially private frequent sequence mining via sampling-based candidate pruning," in ICDE, 2015.

[10] S. Ji, W. Li, P. Mittal, X. Hu, and R. A. Beyah, "Secgraph: A uniform and open-source evaluation system for graph data anonymization and de-anonymization." in USENIX Security Symposium, 2015, pp. 303–318.

[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in TCC, 2006.

[12] X. Cheng, S. Su, S. Xu, and Z. Li, "Dp-apriori: A differentially private frequent itemset mining algorithm based on

transaction splitting," Computers & Security, 2015.

[13] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in SIGMOD, 2000.

[14] S. Su, S. Xu, X. Cheng, Z. Li, and F. Yang, "Differentially private frequent itemset mining via transaction splitting," TKDE, 2015.

[15] H. Li, L. Xiong, and X. Jiang, "Differentially private synthesization of multi-dimensional data using copula functions," in EDBT, 2014.

N.asha Jyothi is a student of, Kakinada Institute of Engineering & Tech., Coringa, East Godavari Dist, AP. Presently she is pursuing his M.Tech [Software Engineering] from this college and he received his B.Tech from Pragathi Engineering College, Surampalem affiliated to JNT University, Kakinada in the year 2015. Her area of interest includes Computer Networks and Object oriented Programming languages, all current trends and techniques in Computer Science.

Mr. D. Srinivas B.Tech., M.Tech is associate professor in Kiet Engineering College. He has 10 years of teaching experience. His area of interest includes Data mining, Networking, Bioinformatics and data science