ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Privacy Preserving of Data Sharing in Social Network Using Mergebysplit

S.Sathiya*, C.Grace Padma

* Research Scholar in M.Phil Computer Science, RVS College of Arts & Science, Sulur, Coimbatore – 641 402, Tamil Nadu, India

Research Guide in M.Phil Computer Science, Associate Professor & HOD, Department of Computer Applications (MCA), RVS College of Arts & Science, Sulur, Coimbatore – 641 402, Tamil Nadu, India

Abstracts

As an expanding amount of social networking data is published and shared for marketable and research purposes, privacy issues regarding the individuals in social networks have become serious concerns. A social network is a social arrangement made up of a set of social actors (such as individuals or organizations) and a set of the dyadic ties between these actors. The social network point of view provides a set of methods for analyzing the structure of whole social entities as well as a range of theories explaining the patterns observed in these structures. The main issues in social networks are privacy. Our existing system provides privacy for social networks, but it more complex, because it uses multiple algorithms. So propose a new enhancement, which is combine two algorithms, and create a new algorithm.

Keywords: Social network, privacy, anonymization.

Introduction

In a social network, individuals are represented by the social activities between individuals are summarized by boundaries. In light of the identification of the effectiveness of information in social networking data for business and research purposes, more and more social networking data have been published and shared in recent years. This, however, raises serious privacy concerns for the individuals whose personal information is contained in social networking data. Vertex identification, where malicious attackers utilize their background knowledge to associate an individual with a specific vertex in published social networking data, is one of the most important privacy issues that have emerged in recent year [15], [16]. [10] First showed that as long as an attacker knows a piece of information about an individual, it is insufficient to protect privacy by only removing the vertex identities. Liu and Terzi [11] later proposed k-degree anonymity that guarantees the privacy protection against degree information. With the explosive growth of information from social networking applications, privacy concerns in releasing social networking data become increasingly important. Various issues, such as vertex identification and link identification, have drawn extensive research interests [15], [16]. Vertex identification [11], [12], [13], [14], [20]finds the one-to-one correspondence of each individual and each vertex in a social network to extract sensitive personal information, and many anonymization and generalization approaches for resisting vertex identification have been introduced in Section 1. This contrasts with link identification [17], [18], [19], [20], which discloses the sensitive relationship between two individuals. To resolve this issue, perturbation [18] with edge addition, edge deletion, and edge swap is proposed. To further address different privacy requirements, edges are classified into multiple types of sensitivities and removed with different priorities [20]. Alternatively a generalization technique is another approach. Hay et al. [13] were able to hide privacy details about each individual by grouping a set of vertices into a super-vertex and inferring the relationships between super-vertices from super-edges. In this paper mainly focusing the privacy issues in social network, everyone knows personal information's when they share their status in public and also we can't share any documents over the social networks.

Literature survey

Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography

Description: In a social network, nodes correspond to people or other social entities, and edges correspond to social links between them. In an effort to preserve privacy, the practice of anonymization replaces names with meaningless unique identifiers. We describe a family of attacks such that even from a single anonymized copy of a social network, it is possible for

an adversary to learn whether edges exist or not between specific targeted pairs of nodes. [1]
Authors: Lars Backstrom, Cynthia Dwork & Jon Kleinberg

An algorithm for parametric community's detection in networks
Description: Anupam Gupta_ Aaron Rothy Jonathan UllmanModularity maximization is extensively used to detect communities in complex networks. It has been shown however that this method suffers from a resolution limit: small communities may be undetectable in the presence of larger ones even if they are very dense. To alleviate this defect, various modifications of the modularity function have been proposed as well as multi-resolution methods. In this paper we systematically study a simple model (first proposed by Pons and lately with a single parameter which balances the fraction of within community edges and the expected fraction of edges according to the configuration model. An exact algorithm is proposed to find optimal solutions for all values of _ as well as the corresponding successive intervals of _ values for which they are optimal. This algorithm relies upon a routine for exact modularity maximization and is limited to moderate size instances. An agglomerative hierarchical heuristic is therefore proposed to address parametric modularity detection in large networks. At each iteration the smallest value of for which it is worthwhile to merge two communities of the current partition is found. Then merging is performed and the data updated accordingly. An implementation is proposed with the same time and space complexity as the well-known Clauset Newman Moore heuristic (CNM) Experimental results on artificial and real world problems show that (i) communities are de-tected by both exact and heuristic methods for all values of the parameter _; (ii) the dendrogram summarizing the results of the heuristic method provides a useful tool for substantive analysis, as illustrated particularly on *Les Mis´erables* data set; (iii) the difference between the parametric modularity values given by the exact method and those given by the heuristic is moderate; (iv) the heuristic version of the proposed parametric method, viewed as a modularity maximization tool, gives better results than the CNM heuristic for large instances.[2]
Authors: Andrea Bettinelli, Pierre Hansen & Leo Liberti

R-MAT: A Recursive Model for Graph Mining
Description: How does a `normal' computer (or social) network look like? How can we spot `abnormal' sub-networks in the Internet, or web graph? The answer to such questions is vital for outlier detection (terrorist networks, or illegal money-laundering rings), forecasting, and simulations. The heart of the problem is finding the properties of real graphs that seem to persist over multiple disciplines. We list such \laws" and, more importantly, we propose a simple, parsimonious model, the \recursive matrix" (R-MAT) model, which can quickly generate realistic graphs, capturing the essence of each graph in only a few parameters. Contrary to existing generators, our model can trivially generate weighted, directed and bipartite graphs; it subsumes the celebrated model as a special case; it can match the power law behaviors, as well as the deviations from them. We present results on multiple, large real graphs, where we show that our parameter algorithm them very well.[3]
Authors: Deepayan Chakrabarti, Yiping Zhan & Christos Faloutsos

Toward Privacy in Public Databases
Description: We initiate a theoretical study of the census problem. Informally, in a census individual respondents give private information to a trusted party (the census bureau), who publishes a sanitized version of the data. There are two fundamentally conflicting requirements: privacy for the respondents and utility of the sanitized data. Unlike in the study of secure function evaluation, in which privacy is preserved to the extent possible given a specific functionality goal, in the census problem privacy, is paramount; intuitively, things that cannot be learned "safely" should not be learned at all. An important contribution of this work is a definition of privacy (and privacy compromise) for statistical databases, together with a method for describing and comparing the privacy offered by specific sanitization techniques. We obtain several privacy results using two different sanitization techniques, and then show how to combine them via cross training. We also obtain two utility results involving clustering.[4]
Authors: Shuchi Chawla, Cynthia Dwork, Frank McSherry, Adam Smith, and Hoeteck Wee4

K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks
Description: Serious concerns on privacy protection in social networks have been raised in recent years; however, research in this area is still in its infancy. The problem is challenging due to the diversity and complexity of graph data, on which an adversary can use many types of background knowledge to conduct an attack. One popular type of attacks as studied by pioneer work is the use of embedding sub graphs. We follow this line of work and identify two realistic targets of attacks, namely, Node Info and Link Info. Our investigations show that *k*-isomorphism, or anonymization by forming

k pair wise isomorphic sub graphs, is both sufficient and necessary for the protection. The problem is shown to be NP-hard. We devise a number of techniques to enhance the anonymization efficiency while retaining the data utility. The satisfactory performance on a number of real datasets, including HEP-Th, EUemail and LiveJournal, illustrates that the high symmetry of social networks is very helpful in mitigating the difficulty of the problem.[5]

Authors: James Cheng, Ada Wai-Chee Fu & Jia Liu

Finding community structure in very large networks

Description: The discovery and analysis of community structure in networks is a topic of considerable recent interest within the physics community, but most methods proposed so far are unsuitable for very large networks because of their computational cost. Here we present a hierarchical agglomeration algorithm for detecting community structure which is faster than many competing algorithms: its running time on a network with n vertices and m edges is O(md log n) where d is the depth of the dendrogram describing the community structure. Many real-world networks are sparse and hierarchical, with m _ n and d _ log n, in which case our algorithm runs in essentially linear time, O(n log2 n). As an example of the application of this algorithm we use it to analyze a network of items for sale on the web-site of a large online retailer, items in the network being linked if they are frequently purchased by the same buyer. The network has more than 400 000 vertices and 2 million edges. We show that our algorithm can extract meaningful communities from this network, revealing large-scale patterns present in the purchasing habits of customers.[6]

Authors: Aaron Clauset,1 M. E. J. Newman,2 and Christopher Moore

Differential Privacy: A Survey of Results

Description: Over the past five years a new approach to privacy-preserving data analysis has borne fruit. This approach differs from much (but not all!) of the related literature in the statistics, databases, theory, and cryptography communities, in that a formal and *ad omnia* privacy guarantee is defined, and the data analysis techniques presented are rigorously proved to satisfy the guarantee. The key privacy guarantee that has emerged is *differential privacy*. Roughly speaking, this ensures that (almost and quantifiably) no risk is incurred by joining a statistical database. In this survey, we recall the definition of differential privacy and two basic techniques for achieving it. We then show some interesting applications of these techniques, presenting algorithms for three specific tasks and three general results on differentially private learning [7].

Authors: Cynthia Dwork

Privacy-Preserving Data Publishing: A Survey of Recent Developments

Description: The collection of digital information by governments, corporations, and individuals has created tremendous opportunities for knowledge and information-based decision making. Driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for the exchange and publication of data among various parties. Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy. The current practice in data publishing relies mainly on policies and guidelines as to what types of data can be published and on agreements on the use of published data. This approach alone may lead to excessive data distortion or insufficient protection. *Privacy-preserving data publishing* (PPDP) provides methods and tools for publishing useful information while preserving data privacy. Recently, PPDP has received considerable attention in research communities, and many approaches have been proposed for different data publishing scenarios. In this survey, we will systematically summarize and evaluate different approaches to PPDP, study the challenges in practical data publishing, clarify the differences and requirements that distinguish PPDP from other related problems, and propose future research directions [8].

Authors: Benjamin c. M. Fung, Ke wang, Rui chen & Philip s. Yu

Iterative Constructions and Private Data Release

Description: In this paper we study the problem of approximately releasing the cut function of a graph while preserving deferential privacy, and give new algorithms (and new analyses of existing algorithms) in both the interactive and non-interactive settings. Our algorithms in the interactive setting are achieved by revisiting the problem of releasing deferentially private, approximate answers to a large number of queries on a database. We show that several algorithms for this problem fall into the same basic framework, and are based on the existence of objects which we call iterative database construction algorithms. We give a new generic framework in which new IDC algorithms give raise to new interactive private query release mechanisms. Our modular analysis simples and tightens the analysis of previous algorithms, leading to improved bounds. We then give a new IDC algorithm (and therefore a new private, interactive query release mechanism) based on the Frieze/Kannan low-rank matrix decomposition. This new release mechanism gives an improvement on prior work in a range of parameters where the size of the database is comparable to the size of the data universe (such as releasing all cut queries on dense graphs). We also give a non-interactive algorithm for efficiently releasing private synthetic data for graph cuts with error O(jV j1:5). Our algorithm is based on randomized response and a no private implementation of the SDP-based, constant-factor approximation algorithm for cut-norm due to Alon and Naor. Finally, we give a reduction based on the IDC framework showing that an efficient, private algorithm for computing sufficiently accurate rank-1 matrix approximations would lead to an improved efficient algorithm for releasing private synthetic data for graph cuts. We leave finding such an algorithm as our main open problem [9].

Authors: Anupam Gupta_ Aaron Roth & Jonathan Ullman

Problem formulations

A social network is a website on the Internet that brings people together in a central location to talk, share ideas and interests, or make new friends. This type of collaboration and sharing of data is often referred to as social media. Unlike traditional media that is often created by no more than 10 people, social media sites contain content that has been created by hundreds or even millions of different people.

Some are the issues in Social Networks

- Privacy
    - Status Leak
    - Images Misuse
    - Personal Information's Leak
    - Database Hacking
- Bad Comments
- Everyone one know a personal information's when they share their status in public.
- We can't share any documents over the social networks

Existing system

Privacy is always a crucial factor in releasing or exchanging data. In the past decade, issues on privacy-preserving data publishing on transaction data, such as record linkage, sensitive attribute linkage, and table linkage, have attracted extensive research interest Record linkage refers to the identification of a record's owner, and its corresponding privacy model, k-anonymity, prevents record linkage by ensuring that at least k records share the same quasi identifier. These existing techniques are given low performance and less security.

Disadvantages

- Complex Structure
- Less Security
- Low Performance

Existing Algorithm

- EdgeConnect
- Createbysplit

Proposed system

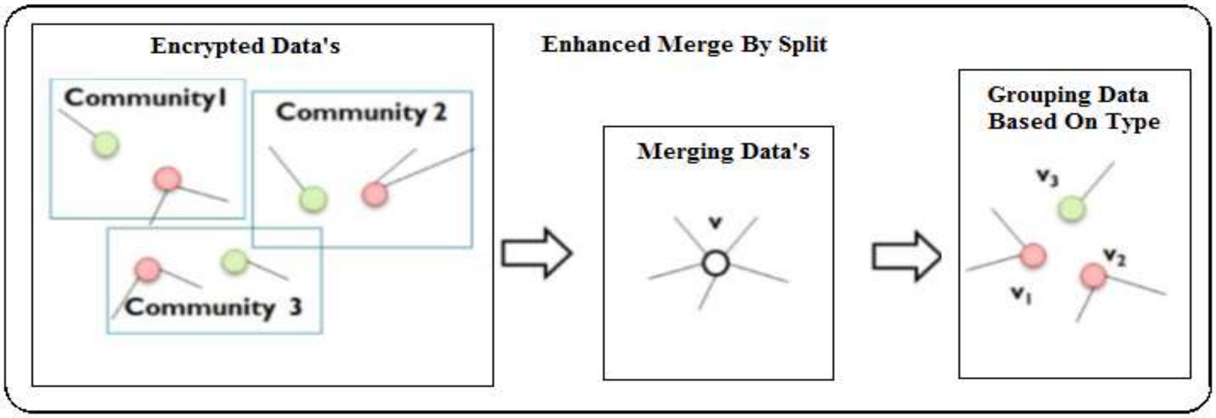
- We propose MergeBySplit Algorithm for the social networks that are difficult to be anonymized with respect to a high privacy level k. In Create by Split, even though Splitting Vertex can generate vertices to increase the possibility of anonymization for the social networks, the algorithm still cannot guarantee finding the solution of every instance of k-SDA. In contrast, MergeBySplit can anonymize every social network, even for the most difficult one.
- We are going to enhance a MergeBySplit algorithm in out proposed system. That is we add privacy enhancement in MergeBySplit.
- In my proposed work, we are using cryptography technique for privacy. Next I reject user details, when they post their status in public. In additional, adding two more features. Such as,
- Story Creations and sharing privately
- Documents Uploading and sharing to another person's privately

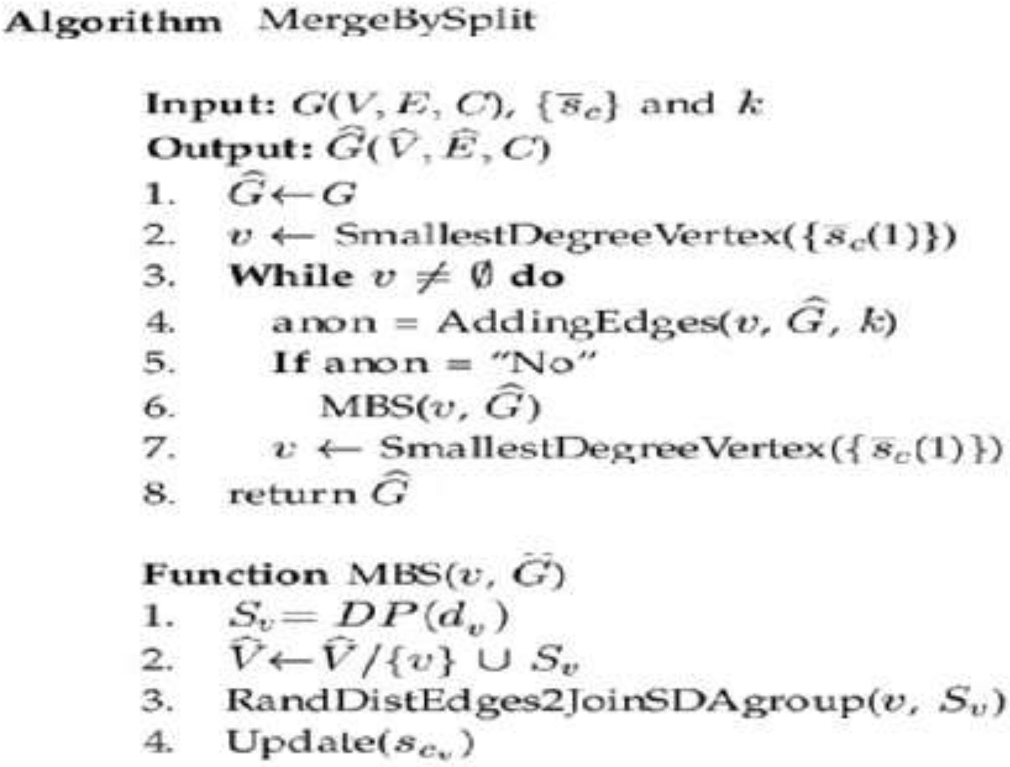
Technique in Proposed System:

- Enhancing MergeBySplit

Enhanced MergeBySplit Algorithm

In this system, we proposed Enhanced Merge by Split algorithm for Clustering, Splitting and privacy. MergebySplit anonymizes the vertices one-by-one in the increasing order of the degrees, and performs Splitting Vertex by allowing each vertex v to be split into more than two substitute vertices protected by the existing k-SDA groups. The rationale of this algorithm is that the creation of k-SDA groups with small degrees allows us to protect any vertex v by splitting v into many cohorts of the generated k-SDA groups. In the worst case, we can split a vertex v of degree dv into dv substitute vertices of degree 1to achieve the anonymization for an arbitrary k, 1<=_ k <=|C|.



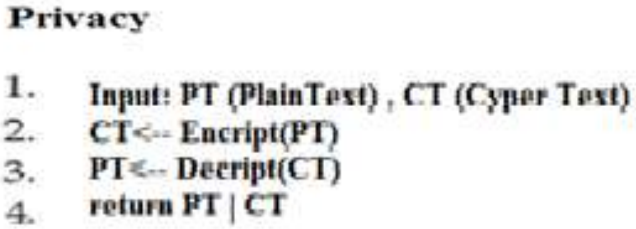


Algorithm MergeBySplit

```
Input: $G(V, E, C)$, $\{\overline{s}_c\}$ and $k$
Output: $\widehat{G}(\widehat{V}, \widehat{E}, C)$
$\widehat{G} \leftarrow G$
$v \leftarrow$ SmallestDegreeVertex($\{\overline{s}_c(1)\}$)
While $v \neq \emptyset$ do
    anon = AddingEdges($v$, $\widehat{G}$, $k$)
    If anon = "No"
        MBS($v$, $\widehat{G}$)
    $v \leftarrow$ SmallestDegreeVertex($\{\overline{s}_c(1)\}$)
return $\widehat{G}$
```

Function MBS(v, $\widehat{G}$)

```
$S_v = DP(d_v)$
$\widehat{V} \leftarrow \widehat{V} / \{v\} \cup S_v$
RandDistEdges2JoinSDAgroup($v$, $S_v$)
Update($s_{c_v}$)
```



Privacy

```
Input: PT (PlainText) , CT (Cyper Text)
CT<-- Encript(PT)
PT<-- Decript(CT)
return PT | CT
```

Result

Authentication

- Authentication is the act of confirming the truth of an attribute of a datum or entity. This might involve confirming the identity of a person or software program, tracing the origins of an artifact, or ensuring that a product is what it's packaging and labeling claims to be. Authentication often involves verifying the validity of at least one form of identification. We authenticate our system by using username and password. The username-password authentication flow can be used to authenticate when the consumer already has the user's credentials.

Application

- Our proposed system we create application for social networks. In this application we share our status and files over the network.

Status Upload

- When upload our status in our network that status stored in to database with encrypted format.

Algorithm Implementation

- Our proposed we implement MergyBySplit with privacy algorithm. This algorithm used for splitting, merging & privacy for status & file uploading.

Log Maintenance

- Log maintenance is used to store the all user details in database with encrypted format.

Processes done in our proposed system

- Merging
- Splitting
- Privacy

Advantages of our proposed system

- Simple Structure
- Combining MergeBySplit & Privacy Algorithm
- High Security

Applications

- Mail Services
- Uploading Websites
- Organizations
- Governments
- Colleges

Conclusion

In this system, we addressed a new privacy issue, community identification, and formulated the problem to protect the community identity of each individual in published social networks. We proposed an Enhanced Merge by Split to find optimal solutions, and also devised scalable results. That is, we enhance a merge by split algorithm for implementing data mining, clustering and cryptography technique in our system. Our system, the added advantages are story creation and sharing to our friend or other. When we create a story and share to a particular person mean, that person having read and write capability for edit and update our story. We can upload file with securely, and also we can download this file.

Future enhancement

For future we implement more features in our system. Our system when a person create a story and share to a particular person mean, that person only having read and write capability that story. In future group of user can enhance a merge by split algorithm in performance basis and also we create a trusted network for users in cloud.

References

1. Lars BackstromDept. of Computer ScienceCornell University, Ithaca NYlars@cs.cornell.edu Cynthia DworkMicrosoft Researchdwork@microsoft.com Jon Kleinberg Dept. of Computer Science Cornell University, Ithaca NY kleinber@cs.cornell.eduWherefore Art Thou R3579X? "Anonymized Social Networks, Hidden Patterns," and Structural Steganography ' Proc. 16th Int'l Conf. World Wide Web (WWW '07), 2007.
2. Andrea Bettinelli_DTI, Universit`a degli Studi di Milano, via Bramante 65, Crema, Italy Pierre Hansen†GERAD, HEC, Montr´eal, Canada Leo Liberti‡LIX, ´Ecole Polytechnique, F-91128 Palaiseau, "An algorithm for parametric communities detection in networks" France (Dated: February 3, 2012) (Dated: February 3, 2012)
3. Deepayan Chakrabarti,Yiping Zhany Christos, Faloutsosz R-MAT: "A Recursive Model for Graph Mining" ACM Computing Surveys, Vol. 42, No. 4, Article 14, Publication date: June 2010.
4. Shuchi Chawla1, Cynthia Dwork2, Frank McSherry2, Adam Smith3 ??, andHoeteck Wee4 1 Carnegie Mellon University, shuchi@cs.cmu.edu, 2 Microsoft Research SVC, {dwork,mcsherry}@microsoft.com, 3 Weizmann Institute of Science, adam.smith@weizmann.ac.il ,4 University of California, Berkeley, hoeteck@cs.berkeley.edu In Memoriam Larry Joseph Stockmeyer 1948–

2004Toward Privacy in Public Databases 2000–2013

5. James Cheng School of Computer Engineering Nanyang Technological University, Singapore jcheng@acm.org ,Ada Wai-Chee Fu Department of Computer Science and Engineering The Chinese University of Hong Kong adafu@cse.cuhk.edu.hk ,Jia Liu Department of Computer Science and Engineering The Chinese University of Hong Kong jliu@cse.cuhk.edu.hk. "K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks" *SIGMOD'10,* June 6–11, 2010, Indianapolis, Indiana, USA. Copyright 2010.
6. M. E. J. Newman, Detecting community structure in networks. *Eur. Phys. J. B* 38, 321–330 (2011).
7. Cynthia D work Microsoft Research dwork@microsoft.com "Differential Privacy: A Survey of Results" 2008.
8. BENJAMIN C. M. FUNGConcordia University, MontrealKE WANG Simon Fraser University, BurnabyRUI CHEN Concordia University, Montreal and PHILIP S. YU University of Illinois at Chicago "Privacy-Preserving Data Publishing: A Survey of Recent Developments" ACM Computing Surveys, Vol. 42, No. 4, Article 14, Publication date: June 2010.
9. Anupam Gupta, Aaron Roth, Jonathan Ullman, "Iterative Constructions and Private Data Release" September 6, 2011.
10. L. Backstrom, C. Dwork, and J.M. Kleinberg, "Wherefore Art Thou r3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography," Proc. 16th Int'l Conf. World Wide Web (WWW '07), 2007.
11. K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2008.
12. J. Li, Y. Tao, and X. Xiao, "Preservation of Proximity Privacy inPublishing Numerical Sensitive Data," Proc. ACM SIGMOD Int'lConf. Management of Data, 2008.
13. M. Hay, G. Miklau, D. Jensen, D.F. Towsley, and P. Weis, "Resisting Structural Re-Identification in Anonymized Social Networks," Proc. VLDB Endowment, vol. 1, pp. 102-114, 2008.
14. B. Zhou and J. Pei, "Preserving Privacy in Social Networks against Neighborhood Attacks," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE '08), 2008.
15. X. Wu, X. Ying, K. Liu, and L. Chen, "A Survey of Privacy- Preservation of Graphs and Social Networks". Springer, 2010.
16. B. Zhou, J. Pei, and W. Luk, "A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data," SIGKDD Explorations, vol. 10, no. 2, pp. 12-22, 2008.
17. J. Cheng, A.W. Fu, and J. Liu, "K-Isomorphism: Privacy PreservingNetwork Publication against Structural Attacks," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2010.
18. X. Ying and X. Wu, "Randomizing Social Networks: A Spectrum Preserving Approach," Proc. SIAM Int'l Conf. Data Mining (SDM '08), 2008.
19. L. Zhang and W. Zhang, "Edge Anonymity in Social Network Graphs," Proc. Int'l Conf. Computational Science and Technology (CSE '09), 2009.
20. M.E. Nergiz, C. Clifton, and A.E. Nergiz, "Multirelational k- Anonymity," IEEE Trans. Knowledge & Data Eng., vol. 21, no. 8, pp. 1104-1117, Aug. 2009.