A Perturbation Based approach for Privacy Preserving Publication of Social Network Graphs

Anirban Chakraborty
National Institute of Technology
Karnataka, Surathkal
anidon.anirban@gmail.com

Annappa B.
National Institute of Technology
Karnataka, Surathkal
annappa@gmail.com

ABSTRACT
Preserving privacy while publishing social network data has become a serious issue with the rapid growth of Social Networks. In this work, we propose a perturbation based approach for privacy preserving publication of social network graphs and evaluate the utility aspect of our proposed method using real world dataset.

Keywords
Randomization, Privacy Disclosure, Social Network

1. INTRODUCTION
Social networks have received dramatic interest in research and development Social networks often contain some private attribute information about individuals as well as their sensitive relationships between different individuals. As more and more social network data has been made publicly available and analyzed in one way or another, privacy preserving publishing of social network data has become an important concern.

2. MOTIVATION
Social networks model social relationships with a graph structure using nodes and edges, where nodes model individual social actors in a network, and edges model relationships between social actors. The attributes of social actors and the relationships between social actors are often private, and directly outsourcing the social networks to a third party may lead to unacceptable disclosures. Privacy preserving data publishing and analyzing techniques on relational data have been well developed. However, the research regarding privacy preservation techniques on social network data is still in its infancy [2].

3. PROPOSED METHOD
Generalization/Clustering, K-anonymity Privacy Preservation and Edge-Randomization are the some popular approaches that exist in literature [2][3][4]. Our method focuses mainly on edge-randomization approach. Mittal et al. [3] gave a perturbation based approach which provides link privacy by modifying the graph structure. We have modified the aforesaid algorithm by incorporating clustering techniques. In this approach the Network Graph is first divided into several clusters and then the perturbation technique is applied. This clustering based approach intends to preserve the local structural property of the graphs. The clustering method adopted for this algorithm is multilevel-clustering, which is fast and easy to be incorporated for network graphs. For testing the utility aspect of the new algorithm reciprocal of the mean distance between the transition probability matrices of the original (G) and the perturbed graph (G') is used.

$$Utility(G, G', L) = 1/(\sum v \in V\ distance(P_v^L(G), P_v^L(G'))/|V|)$$

Here P_v^L denotes the v-th row of the matrix P^L. In the above definition, the parameter L is linked to higher level applications that leverage social graphs [3]. To preserve the local community characteristics, the value of L in this experiment is kept as 3. The distance calculated here is Helinger Distance. The input parameters to the algorithm Perturb(G,t,M) are

i) Original Graph G

ii) Perturbation Parameter t, the length of the random walk

iii) Maximum Loop Count M

Perturb(G,t,M):

```
G' = null;
//The nodes in G are divided into clusters C1,C2,....Ck
for every vertex u in G
   let count = 1;
   for every neighbor v of vertex u
      let loop = 1;
      do
         perform t − 1 hop random walk from vertex v; let z
         denotes the terminal vertex of the random walk;
         loop + +;
      while (u = z or (u, z) ∈ G' ) and (loop ≤ M ));
      if(loop ≤ M and u,z ∈ Cx )
       if count = 1
         add edge (u, z) in G'
       else
         let deg(u) denote degree of u in G; add edge (u, z) in G'
         with  probability (0.5×deg(u)−1)/(deg(u)−1) ;
         count + +;
return G' ;
```

This algorithm is further modified by introducing the notion of significant clusters. We define a cluster C_i to be significant if

$$Ci.size \geq Total\ Number\ of\ Nodes\ /(2 \times No.of\ Clusters)$$

So, according to the algorithm Perturb(G,t,M) if the vertex u lies in a cluster of significant size then only the checking of whether u and z remain in the same cluster or not will be done. The reason behind this is if the vertex u lies in a cluster of a very small size having a few vertices then forcing the perturbed edge to remain in same cluster will be meaningless and may sacrifice purpose of privacy of the whole process.

4. RESULTS

The dataset used for the experiment is crawled from the Facebook Friendship Network of certain individuals. The Social Graph used in this case contains 650+ nodes and 8000+ edges.

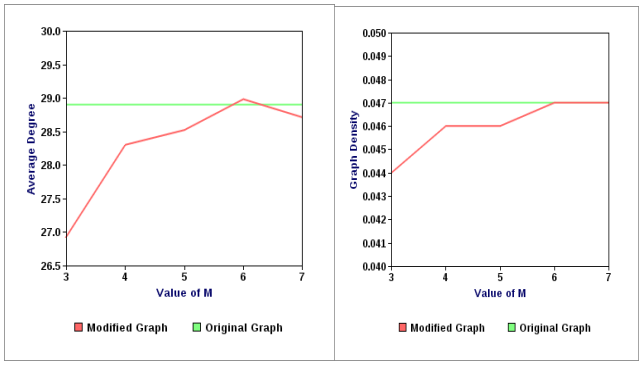

Figure 1: Comparison of Original (G) and Perturbed graph (G') with varying values of M.

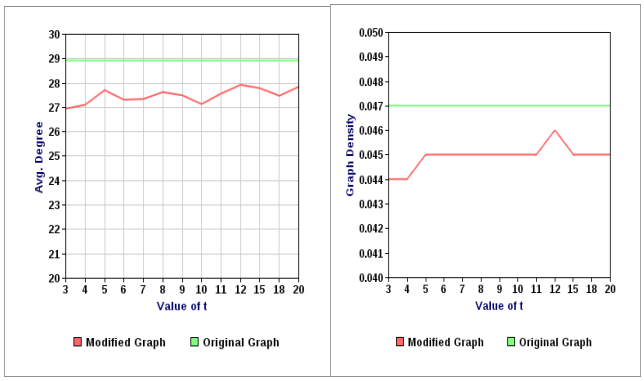

Figure 2: Comparison of Original (G) and Perturbed graph (G') with varying values of t.

So it can be observed that according to the experiments the general graph properties of the modified/perturbed graphs converge to the original one (mostly) at M=7 and t=15 in general.

Figure 3. Shows that Vertex utility of the graph perturbed by the modified algorithm is better than the existing algorithm by 4.5% when t=7 and 6.5% when t=19. However the utility of the new method a slightly lesser than the former one when the value of t is less than 6. Because, as t is the length of the random walk from each node (u according to the algorithm Perturb(G,t,M)), and random walks locally explore the neighborhood of a vertex and with reasonably high probability of staying inside a cluster rather than transitioning to a different cluster [7]. So for smaller values of t performing clustering becomes unnecessary as the random walks will be however terminated inside local clusters only.

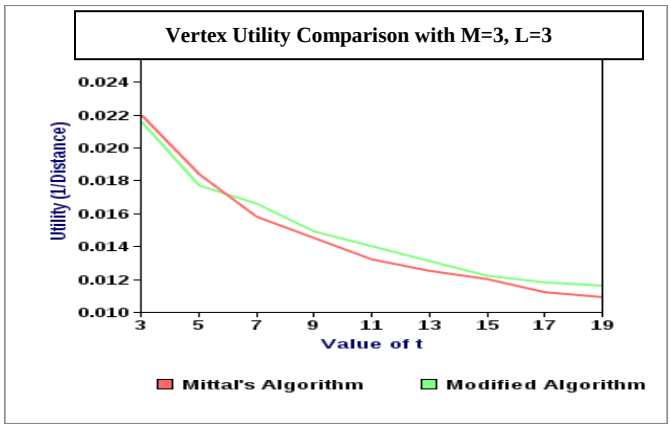

Figure 3: Comparison of Vertex Utility of the modified and existing algorithm.

5. CONCLUSION AND FUTURE WORK

In this work, a perturbation algorithm is proposed that introduces noise in the graph while preserving the graph theoretic properties and local structure of the original graph. Furthermore the utility aspect of the proposed method is also analyzed and comparison of the same is done with the existing algorithm. The benefits of our edge-randomization based approach should also be compared and contrasted with the Clustering and K-Anonymization based approaches. However by protecting the privacy of trust relationships of social graph this perturbation mechanism can be used for real world deployment of systems that leverage social links. The future goals of our research will be,

i) to check the privacy aspects of the proposed model

ii) further evaluation of utility by introducing novel metrics

iii) enhancing the approach to allow dynamic-network analysis

6. REFERENCES

[1] X. Wu, X. Ying and K. Liu 2010. A Survey of Privacy-Preservation of Graphs and Social Networks. In *Managing and Mining Graph Data, Springer Science+Business Media*. 421-453.

[2] B. Zhou, J. Pei and W. Luk 2008. A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data. In *ACM SIGKDD Explorations Newsletter*, 10, 2 (2008), 12-22.

[3] P. Mittal, C. Papamanthou and D. Song 2013. Preserving Link Privacy in Social Network Based Systems. In *Proceedings of NDSS Symposium* (2013).

[4] G. Wang,Q. Liu,F. Li, S. Yang, J. Wu 2013. Outsourcing privacy-preserving social networks to a cloud. In Proceedings of IEEE INFOCOM (2013).

[5]] B. Zhou, J. Pei 2008. Preserving Privacy in Social Networks against Neighborhood Attacks. In *Proceedings of ICDE* (2008) ,