Do Anonymized Social Networks Preserve Communities?

Alina Campan*, Anca Andreica‡ and Traian Marius Truta*

* Department of Computer Science, Northern Kentucky University
Highland Heights, KY 41099, USA
E-mail: (campana1, trutat1)@nku.edu

‡ Department of Computer Science, Babeş-Bolyai University
Cluj-Napoca, Romania
E-mail: anca@cs.ubbcluj.ro

Abstract

Social network anonymization is a powerful privacy-preserving technique for large social networks. In this paper we address the problem of community preservation in anonymized social networks. We investigate how well communities are preserved when a specific anonymization model, namely *k*-anonymity for social network, is used. We use a well-known community detection algorithm based on modularity optimization for our study. To assess the results, we employ three measures (naïve community preservation, community preservation at node level, and normalized mutual information) to quantify community preservation between original and final communities. Our results show that, in general, the community structure is partially preserved in the anonymized networks.

1 Introduction

A significant part of the world population has joined one or more online social networks. As of August 2015, 1,490 million people are active on Facebook, 316 million on Twitter, and 300 million on Google+ [20]. The value of data accumulated by social networks platforms is huge, and they capitalize on marketing opportunities. In many cases, finding people that share similarities in such networks is important, and thus the community detection in social networks has many practical applications. However, there are many situations when the identity of individuals in social networks needs to be protected, and the online social network sites aim to keep the trust of their users. Therefore, they avoid sharing their data with third parties unless the social network data is anonymized and the identity of users protected. This scenario creates a problem for a third party user/organization that needs to extract useful information from the anonymized social network. This paper explores if the communities detected from anonymized social network are similar to the ones that are detected from the original social networks. If they are similar, then mining communities from anonymized networks can be useful.

There are many approaches to anonymize social networks, and they are classified in three categories: anonymization via clustering, anonymization via graph modification, and a hybrid approach [8, 21, 24]. Social network anonymization models and related privacy in social networks issues are covered in [22]. For this community preservation study, we use the *k*-anonymity for social networks [8], a model from the anonymization by clustering family.

Community detection in social networks is a well-studied problem. An overview of community detection algorithms that are most commonly used in social networks can be found in [14]. For this work, we use the Louvain community detection algorithm [3, 19], which is a heuristic algorithm based on a quality function called modularity optimization [18].

Community preservation in anonymized social network is a new topic that is recently investigated in the literature [7, 9, 23]. In our previous work, we proposed two community preservation measures and we ran preliminary experiments [7]. A new approach of anonymizing social networks to improve community preservation is introduced in [9]. A review of the literature on privacy issues in social networks, with a special focus on communities in social networks is presented in [23].

Comparing community preservation between the communities determined in the original network and in the anonymized network, respectively, is not a trivial task. First, the anonymized network (using the *k*-anonymity model for social networks) cannot be used directly for community detection; we employ an additional step called de-anonymization (details will be provided later). Second, after communities are determined both in the original and anonymized networks, we use several measures to quantify community preservation between original and final communities. These measures are: naïve community preservation and community preservation at node level [7], and the normalized mutual information [12] which is the state-of-the art community preservation measure. Third, we compare the results obtained with the community preservation measures listed above.

The main contribution of the paper is the framework of studying the community preservation in the context of anonymized social networks. In addition, this is the first paper that uses normalized mutual information for this problem.

The remaining of this paper is organized as follows. Section 2 presents the *k*-anonymity for social networks model and the RMAT de-anonymization approach that we used to "guess" the original social network structure. Section 3 describes the Louvain community detection algorithm and the three measures to compute the community preservation. Section 4 shows the workflow of our experiments, the description of the social network datasets, and our experimental results. Conclusions are presented in Section 5.

2 *K*-Anonymity for Social Networks and De-anonymization Process

The *k*-anonymity for social networks model is a clustering anonymization approach for social networks in which disjoint groups of at least *k* nodes are clustered together in "super-nodes". The number of original nodes, as well as the number of edges within the super-node, are stored in each super-node. Two super-nodes are connected if at least one pair of nodes from the super-nodes is connected in the original network. Each such edge (called a "super-edge") also stores the number of original edges it replaces. A complete presentation of the *k*-anonymity for social network model, as well as criteria for creating the original nodes grouping can be found in [8].

Figure 1 shows an example of a social network, $\mathcal{G}_1$, and Figure 2 shows one of its possible anonymized networks, $\mathcal{AG}_1$, for $k = 4$.

An anonymized network has a smaller number of nodes than the original network (at most $\lfloor n / k \rfloor$, where n represents the number of nodes in the original network, and $\lfloor x \rfloor$ represents the floor of x). Also, the resulting nodes and edges also have a weight value (number of edges) with no equivalent in the original network. To overcome these differences in our assessment of community preservation, we chose to "de-anonymize" the anonymized network, by trying to reverse the anonymization process and create replicas (or instances) of the original network. Following the de-anonymization process, the same number of nodes and edges will exist in the de-anonymized networks as in the original one, and a fair comparison of communities is possible. The de-anonymization employed in this phase is called R-MAT de-anonymization since it adapts a scale free network generator approach called R-MAT [11] to the de-anonymization problem. Figure 3 shows a possible de-anonymized network for our running example. Full details regarding R-MAT de-anonymization can be found in [1].

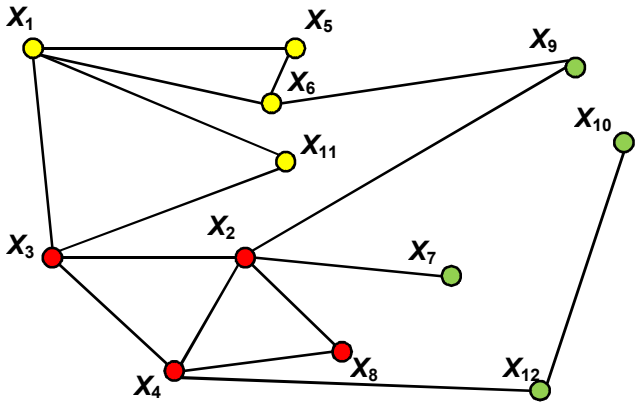


Figure 1: Social network example, $\mathcal{G}_1$.

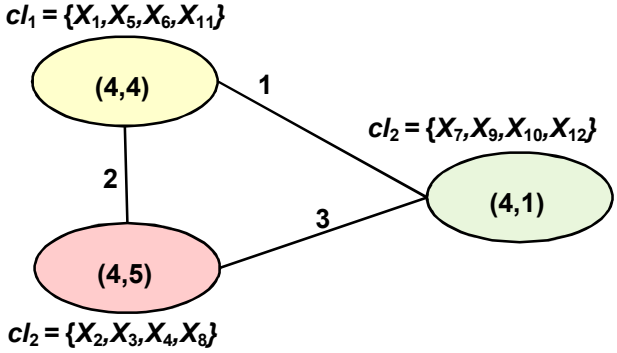


Figure 2: Anonymized social network, $\mathcal{AG}_1$.

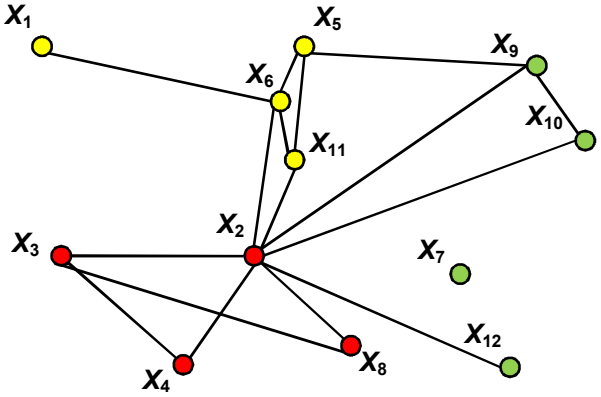


Figure 3: De-anonymized social network, $\mathcal{DG}_1$.

3 Community Preservation

We study community preservation using the Louvain method for community detection [3, 19]. This method is based on maximizing the modularity function [2, 18]. Unfortunately, the optimal solution based on modularity maximization is NP-complete [5], thus the Louvain algorithm tries to approach the optimal solution via a greedy approach.

The modularity is a quality function computed for a graph partitioned in communities. This function is defined for a network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ as follows [18]:

$$Q = \frac{1}{2m} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j),$$

where

- n is the number of nodes in the network ($n = |\mathcal{G}|$);
- m is the number of edges in the network ($m = |\mathcal{E}|$);
- c_i and c_j are the communities to which nodes X_i and X_j have been assigned (all pairs of nodes X_i and X_j are considered);
- A_{ij} represents whether there is an edge between nodes X_i and X_j ($A_{ij} \neq 0$) or not ($A_{ij} = 0$);
- k_i and k_j are the degrees of nodes X_i and X_j;
- $\delta(c_i, c_j)$ is 1 if nodes X_i and X_j belong to the same community ($c_i = c_j$) and 0 otherwise.

Since the terms from the modularity sum are non-zero only for nodes from the same community, the modularity function can be rewritten as [14]:

$$Q = \sum_{c=1}^{n_c} \left[\frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right],$$

where

- n_c is the number of communities;
- l_c is the total number of edges joining nodes from community c (inter-cluster edges);
- d_c is the sum of the degrees of nodes from c.

In the above formula, $\frac{l_c}{m}$ is the actual fraction of edges in the network inside the community and $\left(\frac{d_c}{2m}\right)^2$ is the expected fraction of edges that would be there if the network was a random network with same expected degree for each node.

This modularity function sometimes creates one or more large community that contain very dense small communities that are weakly connected [14]. Thus such a large community can be split into several smaller more compact communities. In order to address this problem, a new modularity function is defined using a resolution parameter r [14]:

$$Q_r = \sum_{c=1}^{n_c} \left[\frac{l_c}{m} - r \left(\frac{d_c}{2m} \right)^2 \right].$$

The larger the resolution parameter, smaller communities are sought for. When r is 1, then Q_r is equivalent to the original modularity function Q.

For our experiments we used the Louvain algorithm [3, 19] for modularity optimization with resolution 1 (the default value). This algorithm performs iteratively two steps. In the first step, each node forms a community and then nodes are moved between communities to maximize modularity gain. In the second step, a weighted network is built from the network obtained at the end of first step, by replacing each community with a node, and intracommunity edges with a weighted edge where the weight represents the number of edges between a pair of communities. The first step is then applied to this new network. The algorithm ends when there is no change in the first step, and thus modularity gain is maximized. Full details regarding this algorithm can be found in [3].

To compare community preservation between the original network and its corresponding de-anonymized versions (see Section 2), we use three measures: *naïve community preservation* [7], *community preservation at node level* [7], and *normalized mutual information measure* [12]. In our previous work [7], we developed the first two community preservation measures, and in this paper we will compare these measures with the normalized mutual information measure, which is a state-of-the art measure for comparing community structure identification [13]. In the remaining of this section, we will briefly describe these three measures.

In the *naive community preservation* (NCP) measure, we count how many nodes from the original communities are preserved in the same community after the processes of anonymization and de-anonymization.

In the *community preservation at node level* (NCPL) we consider community preservation for each node, individually. For each node X_i, we label the corresponding communities in original network and anonymized network as $C(X_i)$ and $C'(X_i)$, respectively. We define the community preservation for the node X_i as the number of nodes that are common between $C(X_i)$ and $C'(X_i)$ divided to the number of nodes that belong to at least one of these two communities. In other words,

$$\mathrm{CPNL}(\mathrm{X_i}) = \frac{|\mathrm{C}(\mathrm{X_i}) \cap \mathrm{C}'(\mathrm{X_i})|}{|\mathrm{C}(\mathrm{X_i}) \cup \mathrm{C}'(\mathrm{X_i})|},$$

where $|S|$ represents the cardinality of the set S.

The final community preservation at node level measure is the average of all individual values:

$$\mathrm{CPNL}(\mathrm{G}, \mathrm{G}') = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{CPNL}(\mathrm{X_i})}{\mathrm{n}}.$$

The *normalized mutual information* (NMI) measure is an information-theoretic measure which compares two partitions of a finite set [12, 13]. This measure is based on a confusion matrix, $\mathcal{N}$, where the rows represent communities from the original social network and the columns – communities from the anonymized social network. The element n_{ij} represents the number of nodes from the community i in the original social network that are found in the community j in the anonymized social network. This measure is computed using the following formula:

$$\mathrm{NMI}(\mathrm{G},\mathrm{G}') = \frac{-2\sum_{i=1}^{C_G}\sum_{j=1}^{C_{G'}} n_{ij}\cdot\log\left(n_{ij}\cdot n/n_i\cdot n_j\right)}{\sum_{i=1}^{C_G} n_i\cdot\log(n_i/n)+\sum_{j=1}^{C_{G'}} n_j\cdot\log\left(n_j/n\right)}.$$

where:

- C_G is the number of communities in original social network;
- $C_{G'}$ is the number of communities in anonymized social network;
- n is the number of nodes in the network;
- n_i is the sum over row i for the confusion matrix;
- n_j is the sum over column j for the confusion matrix;

A value 1 is obtained for the normalized mutual information when the two community sets are identical.

More details regarding NCP and NCPL measures can be found in [7]. Also, a complete presentation of the NMI measure can be found in [12, 13].

4 Experimental Results

The workflow of our experiments is shown in Figure 3. For each initial social network, we compute its anonymized version for k = 5, 10, 15, 20, 25, and 50. Each such anonymized network is de-anonymized using the R-MAT anonymization process. Since this process involves a random selection, we chose to de-anonymize each network 5 times and the reported results are the average of those multiple de-anonymizations. For the initial social network, as well as for all de-anonymized networks, we compute the communities using the Louvain method. Finally, we compare communities detected in the original and de-anonymized networks using the naïve community preservation, community preservation at node level, and normalized mutual information measures.

We used three publically available datasets and two synthetic datasets in our experiments. They are described below:

- *Enron* dataset is a network of email exchanges [15]. It is an undirected network with 36,692 nodes and 183,831 edges. Each node in this network represents an email address. An edge exists between two nodes if at least one email was sent from one node to the other from that edge.
- *AstroPh* is a Collaboration network of Arxiv Astro Physics [15]. It is an undirected network with 18,772 nodes and 198,110 edges.
- *HepTh* is a Collaboration network of Arxiv High Energy Physics Theory [15]. It is an undirected network with 9,877 nodes and 25,998 edges.
- *ERRandom* is a synthetic graph generated using SNAP library [16]. This generator uses the well-known Erdos-Reiny graph model [4]. *ERRandom* is an undirected network with 29,999 nodes and 150,000 edges.
- *RndPowerLaw* is a syntetic graph generated using SNAP library [16]. This generator uses the configuration model [6, 17], which generates a scale-free network [10]. *RndPowerLaw* is an undirected network with 30,000 nodes and 109,423 edges.

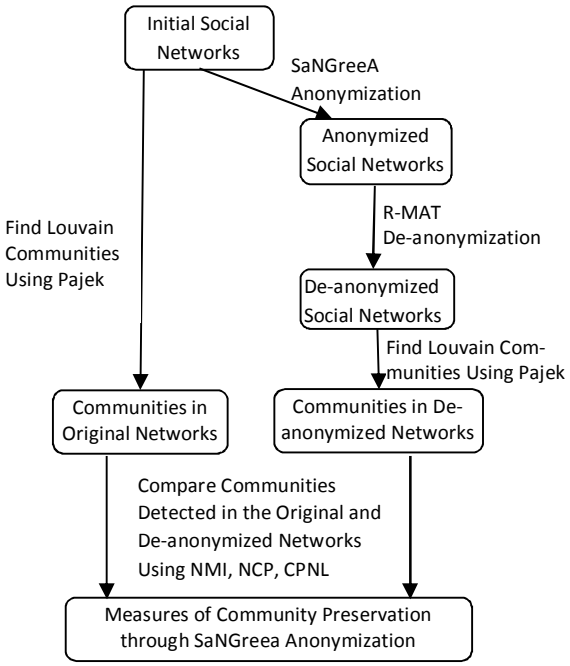


Figure 4: Experiments workflow

To execute our experiments, we implemented SaNGreeA algorithm (used for anonymization), R-MAT De-anonymization, and computation of the NMI, NCP and CPNL measures in Java. The Louvain community detection algorithm was executed using Pajek. All the experiments were performed on an Intel® Core™ i7-2640M 2.80 GHz CPU machine with 8 GB memory running on 64 bit Windows 7 Home Premium.

Figures 5 – 9 show the values for the three community preservations measures for *Enron*, *AstroPh*, *HepTh*, *ERRandom*, and *RndPowerLaw* datasets, respectively. The horizontal axis shows the value of k (from k-anonymity). The last value, *max*, in each of the Figures, represents the value of k equal with the number of nodes in each network. In this extreme case, all nodes are grouped into one supercluster, in other words no structure information is preserved by the anonymization, and the de-anonymized networks are in fact RMAT generated networks of size *max*. We expect that, for smaller values of k, the CPNL, NCP, and NMI measures to have a higher value that in this extreme case. In most cases (except *ERRandom*) this assumption is true; also, as expected, a smaller value of k means better community preservation. We also notice that all three measures have a similar slope, in other words they all agree on the extent of community preservation between two original and de-anonymized networks.

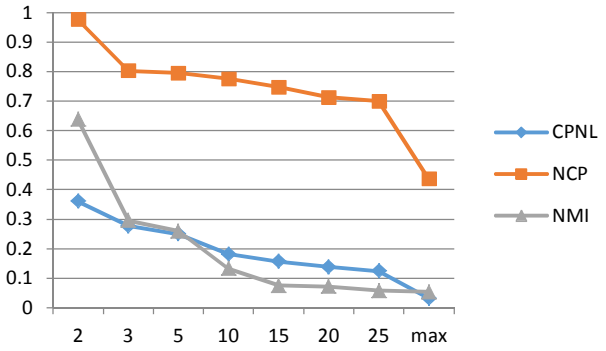


Figure 5: CPNL, NCP, NMI values for *Enron*

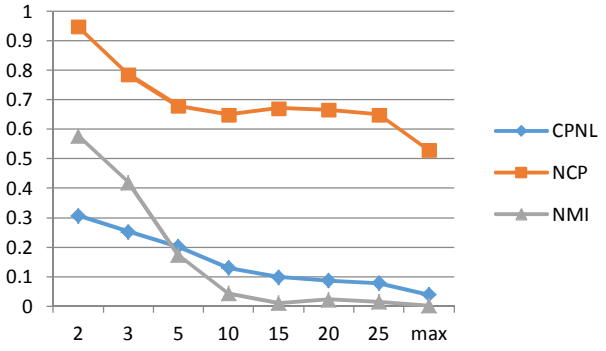


Figure 6: CPNL, NCP, NMI values for *AstroPh*

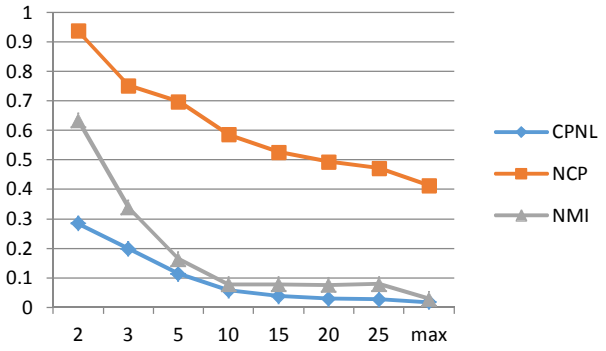


Figure 7: CPNL, NCP, NMI values for *HepTh*

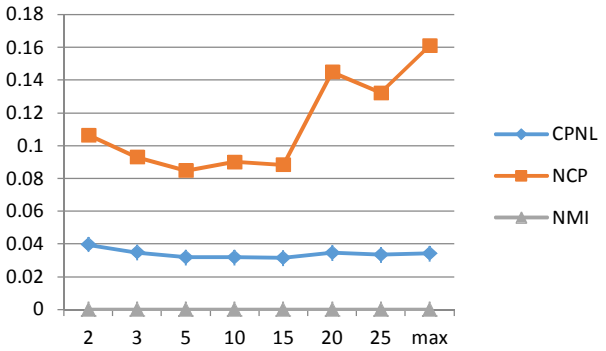


Figure 8: CPNL, NCP, NMI values for *ERRandom*

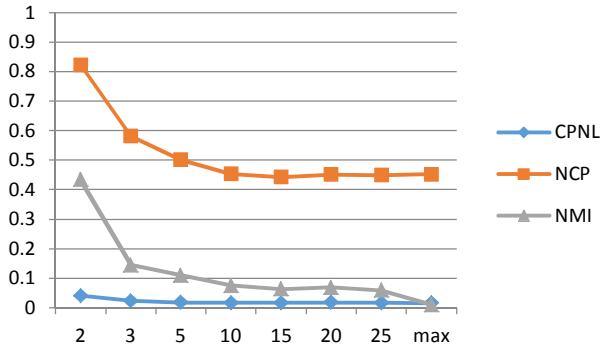


Figure 9: CPNL, NCP, NMI values for *RndPowerLaw*

To illustrate better the difference between a small k value and the extreme case (when k is equal to the size of the network) we also show the improvement factor of those three measures over the extreme case. The improvement factor is the value of a community preservation measure for a given k divided by the value corresponding to the extreme case. Due to space limitation, we show this improvement factor for two datasets only, *Enron* and *ERRandom*, in Tables 1 and 2. For the other three datasets, the improvement factor is consistently above 1 and similar to the *Enron* case (and, in fact, the improvement factors are higher than in the Enron case for *AstroPh*, *HepTh*, and *RndPowerLaw*).

Table 1: Improvement Factor for *Enron*

k	$\frac{CPNL(k)}{CPNL(max)}$	$\frac{NCP(k)}{NCP(max)}$	$\frac{NMI(k)}{NMI(max)}$
2	11.49	2.24	11.82
3	8.82	1.84	5.48
5	7.96	1.82	4.82
10	5.76	1.77	2.46
15	4.98	1.71	1.40
20	4.41	1.63	1.33
25	3.94	1.60	1.08

Table 2: Improvement Factor for *ERRandom*

k	$\frac{CPNL(k)}{CPNL(max)}$	$\frac{NCP(k)}{NCP(max)}$	$\frac{NMI(k)}{NMI(max)}$
2	1.15	0.66	0.01
3	1.02	0.58	0.14
5	0.93	0.53	0.25
10	0.93	0.56	0.58
15	0.92	0.55	0.65
20	1.01	0.90	0.70
25	0.98	0.82	0.80

As seen in both Figure 8 and Table 2, the results for *ERRandom* are, at the first view, surprising. With only a few exceptions, the community preservation is worse than in the random case. This surprising result can be explained based on the RMAT de-anonymization process, which

generates a power-law network in each super node and thus destroys the uniform randomness of the Erdos-Reiny model. Nevertheless, more experiments need to be performed for *ERRandom,* to understand why the slope varies between the three community preservation measures.

5 Conclusions

In this paper, we studied how well communities are preserved when social networks are anonymized. For this purpose, we used three community preservation measures: naïve community preservation, community preservation at node level, and normalized mutual information measure. The RMAT de-anonymization method used after the social networks were anonymized with *SaNGreeA* algorithm (to become *k*-anonymous social networks) is able to preserve, for most networks, the initial communities. As expected, a smaller value of *k* allows better community preservation.

References

[1] Yasmeen Alufaisan and Alina Campan. Preservation of centrality measures in anonymized social networks. *ASE/IEEE International Conference on Privacy, Security, Risk, and Trust (PASSAT 2013)*, 2013.

[2] A. Arenas, J. Duch, A. Fernandez, and S. Gomez. Size Reduction of complex networks preserving modularity. *New Journal of Physics*, 9, 176, 2007.

[3] Vincent D. Blondel, Jean-Loup Guillaume, Renauld Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, 1742-5468, 2008.

[4] Bela Bollobás. Random graphs (2nd ed.). Cambridge University Press. ISBN 0-521-79722-5, 2001.

[5] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On Modularity Clustering, *IEEE Transactions on Knowledge and Data Engineering*, 20(2), 172–188, 2008.

[6] Tom Britton, Maria Deijfen, and Anders Martin-Lof. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124(6): 1377–1397, 2006.

[7] Alina Campan, Yasmeen Alufaisan, and Traian Marius Truta. Preserving communities in anonymized social networks. *Transactions on Data Privacy*, 8(1): 55–87, 2015.

[8] Alina Campan, and Traian Marius Truta. A Clustering Approach for Data and Structural Anonymity in Social Networks. *ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD)*, 2008.

[9] Jordi Casas-Roma and Francois Rousseau. Community-preserving generalization of social networks. *Proceedings of the Social Media and Risk ASONAM 2015 Workshop*, 2015.

[10] Michele Catanzaro, Marian Boguna, and Romualdo Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Physical Review*, E 71, 027103, 2005.

[11] Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos. R-MAT: A recursive model for graph mining. *SIAM International Conference on Data Mining (SDM'04)*, 442–446, 2004.

[12] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiments*, 2005(09): P09008, 2005.

[13] Jan Dreier, Philipp Kuinke, Rafael Przybylsky, Felix Reidl, Peter Rossmanith, and Somnath Sikdar. Overlapping communities in social networks. arXiv preprint arXiv:1412.4973, 2014.

[14] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486 (3–5), 75–174, 2010.

[15] Jure Leskovec and Andrej Krevl. SNAP datasets: Stanford large network dataset collection, `http://snap.stanford.edu/data`, 2014.

[16] Jure Leskovec and Rok Sosic. A general purpose network analysis and graph mining library in C++, `http://snap.stanford.edu/snap`, 2014.

[17] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, Vol. 6, 161–180, 1995.

[18] M. E. J Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, E 69 (2), 026113, 2004.

[19] Randolf Rotta and Andreas Noack. Multilevel Local Search Algorithms for Modularity Clustering. *Journal of Experimental Algorithms*, 6(2.3), 2011.

[20] Statista. Global social networks ranked by number of users 2015. `http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/`, November 2015.

[21] Elena Zheleva and Lise Getoor. Preserving the privacy of sensitive relationships in graph data. *ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD),* 153–171, 2007.

[22] Elena Zheleva, Evimaria Terzi, and Lise Getoor. Privacy in social networks. *Synthesis Lectures on Data Mining Series*, Morgan and Claypool Publishers, 2012.

[23] Bin Zhou. Privacy issues in discovering communities in social networks. G. Paliouras et al. (eds.), *User Community Discovery*, Human–Computer, Chapter 6, 2015.

[24] Bin Zhou and Jian Pei. Preserving privacy in social networks against neighborhood attacks. *IEEE International Conference on Data Engineering (ICDE)*, 506–515, 2008.