Also available at http://amc-journal.eu
ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 7 (2014) 461–477

Triangle randomization for social network data anonymization[*]

Ljiljana Brankovic
University of Newcastle, Callaghan NSW 2308, Australia

Nacho López
Universitat de Lleida, C.Jaume II, 69, E-25001 Lleida, Spain

Mirka Miller
University of Newcastle, Callaghan NSW 2308, Australia
University of West Bohemia, Pilsen, Czech Republic
King's College London, The Strand, United Kingdom

Francesc Sebé
Universitat de Lleida, C.Jaume II, 69, E-25001 Lleida, Spain

Received 25 July 2011, accepted 29 August 2012, published online 27 June 2014

Abstract

In order to protect privacy of social network participants, network graph data should be anonymised prior to its release. Most proposals in the literature aim to achieve k-anonymity under specific assumptions about the background information available to the attacker. Our method is based on randomizing the location of the triangles in the graph. We show that this simple method preserves the main structural parameters of the graph to a high extent, while providing a high re-identification confusion.

Keywords: Anonymity, privacy, social network.

Math. Subj. Class.: 68R10

[*] This paper is a part of Bled'11 Special Issue.
This work was partly supported by: the Spanish Ministry of Economy and Competitiveness through project TIN2010-18978; RGC CEF grant G0189479, The University of Newcastle; and by a Marie Curie International Incoming Fellowship within the 7th European Community Framework Programme.

E-mail addresses: Ljiljana.Brankovic@newcastle.edu.au (Ljiljana Brankovic), nlopez@matematica.udl.cat (Nacho López), mirka.miller@newcastle.edu.au (Mirka Miller), fsebe@matematica.udl.cat (Francesc Sebé)

1 Introduction

A social network comprises a set of participants and the relations among them. Such a network is naturally modelled by a graph structure, where each participant in the network is assigned to a vertex and relations are represented by edges connecting pairs of vertices.

Analysis of social network graph data [15, 17], and evolution [9] is a source of valuable information for numerous research areas, including sociology, psychology, economics and epidemiology, to mention but a few. The results of such research have shed light on wide range of problems, from the dynamics of happiness [5], obesity [4] and smoking [12] to crime investigation [6]. Notwithstanding the benefits that such studies offer in various areas of life, they also introduce threats to individuals' privacy. Social networks contain sensitive personal data whose publication or exchange would compromise their members' privacy. As an example, consider a graph in which nodes are e-mail addresses and the edges represent 'message exchange' relations. The list of people a person communicates with is an example of very sensitive data. In order to alleviate privacy risks, it is generally accepted that access to social network data for scientific research requires a pre-processing phase to reduce the opportunities for inferring information about individuals. This should be done in such a way so as to maintain a high quality of the transformed data so that the analysis can be performed with acceptable accuracy.

Privacy in social networks considers two basic aspects:

- *Vertex anonymity:* It should not be possible to infer the identity of vertices in the published anonymized network.
- *Edge anonymity:* Given two social network participants, it should not be possible to infer whether an edge exists between their corresponding vertices, *i.e.*, whether they are related.

Naive anonymization provided by the removal of individuals' identifiers has been proven insufficient since background knowledge such as the vertex degree or the neighborhood subgraph of some participants often permits the identification of many of the vertices. Therefore, additional privacy measures must be applied. These measures can be classified into:

- *Generalization based techniques:* Nodes and edges are first clustered and then collapsed into supervertices and superedges [2, 8].
- *Perturbation based techniques:* The original social network is modified by adding and/or removing vertices and/or edges [3, 10, 19, 20, 21].

Privacy of any anonymization technique depends on the previous knowledge the attacker is assumed to have. In [10] the attacker is assumed to know the degrees of all the vertices in the network, and thus the original network is modified by edge additions and/or deletions until the degree sequence is k-anonymous [14]. A similar approach is given in [21], which provides k-anonymity even when the attacker has prior knowledge about the neighborhood subgraph of target vertices.

Recently, *information-theoretic models* have been proposed, which seek to achieve robustness against any background structural knowledge of the attacker. The k-Symmetry [19] and k-Automorphism [20] models aim to protect against "identity disclosure" (vertex anonymity) by adding vertices and edges until, for each vertex of the graph, there exist at least $k-1$ other vertices that are structurally equivalent to it. The k-Isomorphism [3] model also considers edge privacy and generates an anonymous graph that consists of k disjoint isomorphic subgraphs. Such models impose hard structural requirements on anonymized graphs that require extensive modifications to the original graph. From the computational point of view, implementing these security models depends on the capability to cope with some known NP-hard problems on graphs.

In [16], n-confusion is proposed as a privacy model that generalizes k-anonymity. This model requires that the set of nodes from the released graph that can correspond to any given identity has a size larger or equal to n. In this paper, the privacy is analyzed from that point of view.

1.1 Our approach

Information-theoretic models achieving k-anonymity against any structural knowlege [3, 19, 20] are hard to put into practice and require extensive perturbation of the original graph. We claim adequate privacy can be achieved by means of simpler random noise techniques.

Structural background knowledge may be of diverse nature. We focus our attention on two specific parameters: vertex degree and the number of triangles passing through a given vertex. In [19] it is shown that their combined knowledge provides a high re-identification power in a trivially anonymized graph. Both parameters are implicit parts of other structural properties. For instance, knowledge of the neighbourhood subgraph of a vertex implies knowing its degree and all triangles passing through the vertex. Any anonymization procedure that perturbs both parameters also impairs re-identification methods based on more complex structural knowledge implicitly involving the parameters. An additional important aspect is that both parameters are easy to measure so that they can be part of computationally efficient anonymization techniques.

We propose a method that first removes the triangles of the graph (by deleting at least one edge of each triangle) and next randomly adds edges to the resulting graph so as to create approximately as many triangles as there were in the original graph. This triangle randomization process makes the information about triangles passing through a vertex less reliable for matching purposes, while it also perturbs the degree of the vertices in the graph. At the same time, the global structure of the graph remains very similar to the original one. It is easy to see that this procedure preserves the connected components since removing one edge from a triangle can never disconnect the graph and new edges are only added between vertices that are connected by a path of length 2. Experimental results show that other graph structure parameters are preserved to a high extent.

1.2 Main contributions

In this paper we propose a novel perturbation technique for preserving privacy in social networks, based on randomization of the locations of triangles in the graph. Our technique (1) is simple and can be efficiently implemented on large graphs; (2) provides high level of privacy, as supported by our experiments; (3) provides a high degree of data utility, as supported by experimental evidence.

The organisation of the paper is as follows. In the next section we present the anonymization algorithm and give bounds on the number of triangles in the perturbed graph. In Section 3 we analyse the privacy of our technique by providing the bounds for the degree and a number of triangles passing through each vertex of the perturbed graph, and using these bounds to measure degree-triangle anonymity. Section 4 presents our experimental results on real and synthetic data sets, both in terms of utility and privacy. In Section 5 we give some concluding remarks and directions for future research.

2 Anonymization procedure

We model a social network as an undirected graph $G = (V, E)$, where V and E are the sets of vertices and edges, respectively. If $u, v, w \in V$ and $uv, vw, wu \in E$, then (u, v, w) is said to be a triangle of G.

Our proposal is a perturbation-based technique consisting of two rounds, where anonymization is achieved by means of edge removal and addition. Its description is given in Algorithm 1.

```
   Input: Original graph $G = (V, E)$
   Output: Masked graph $G$
 1 $T = NumTriangles(G)$;
 2 while $NumTriangles(G) > 0$ do
 3     $(v_i, v_j, v_k) = TakeTriangleAtRandom(G)$;
 4     $b = TakeAtRandomFrom(\{0, 1, 2\})$;
 5     if $b$=0 then
 6         $RemoveEdge(G, (v_j, v_k))$
 7     else if $b$=1 then
 8         $RemoveEdge(G, (v_i, v_k))$
 9     else
10         $RemoveEdge(G, (v_i, v_j))$
11     end
12 end
13 while $NumTriangles(G) < T$ do
14     $(v_i, v_j) = TakeEdgeAtRandom(G)$;
15     $b = TakeAtRandomFrom(\{0, 1\})$;
16     if $b$=0 then
17         $v_k = TakeNeighborAtRandom(G, v_i)$;
18         if $v_k \neq v_j$ and $(v_k, v_j) \notin E$ then
19             $AddEdge(G, (v_k, v_j))$
20         end
21     else
22         $v_k = TakeNeighborAtRandom(G, v_j)$;
23         if $v_k \neq v_i$ and $(v_k, v_i) \notin E$ then
24             $AddEdge(G, (v_k, v_i))$
25         end
26     end
27 end
28 return (G)
```

Algorithm 1: Triangle randomization algorithm

The first round (steps 1-12) is a procedure that randomly selects a triangle of G, then randomly selects one of its edges and removes it; this is repeated until there are no more triangles left. Note that the removal of a single edge may cause the deletion of several triangles. The second round (steps 13-27) adds edges that create one or more triangles each.

Note that the number of triangles after the algorithm has finished, T', may not equal the original number T. Since Algorithm 1 terminates when the addition of a single edge gives a graph with the total number of triangles $T' \geq T$, it is very likely that this results in $T' \approx T$. This fact introduces some additional uncertainty that further obstructs possible matching attempts.

We next provide a bound on the number of triangles T' in the perturbed graph G'.

Proposition 2.1. *Let G be a graph that has been perturbed into G' using Algorithm 1, and let $T \geq 1$ and T' be the number of triangles in G and G', respectively. Then $T' < T + \Delta' - 1$, where Δ' is the maximum degree in G'.*

Proof. Let $G' = (V', E')$ and $G'' = (V'', E'')$, $V'' = V'$ and $E'' = E' \setminus \{uv\}$, where uv is the very last edge added by the Algorithm 1 (informally, G'' is the graph obtained by Algorithm 1 just before the very last edge, say uv, is added), and let T'' be the number of triangles in G''. Then $T' = T'' + T^{uv}$, where T^{uv} is the number triangles created by addition of the edge uv. Then we have $T^{uv} = |N''(u) \cap N''(v)| \leq |N''(u)| < \Delta'$, where $N''(u)$ is the neighbourhood of u in G''. Since $T'' \leq T - 1$ we have $T' < T + \Delta' - 1$. □

The previous proposition states that T' is at most $T + \Delta' - 2$. From the proof, this extremal situation is given when $T'' = T - 1$ and $T^{uv} = |N''(u) \cap N''(v)| = |N''(u)| = \Delta' - 1$. We pose the following open problem.

Problem 2.2. Is there a family of graphs G with arbitrarily large values of T that can be perturbed to graphs G' such that $T' = T + \Delta' - 2$?

3 Data privacy

In this section, we analyse the privacy of our proposal assuming that the attacker's background knowledge comprises the degree and the number of triangles passing through some vertices. The objective of the proposed method is to disrupt any attempt to obtain knowledge about the identity of the nodes in the published anonymized network. So as to evaluate the extent to which this objective is achieved, we need some methods for measuring it. These are the privacy metrics.

3.1 Degree-triangle variation

Let us consider a vertex $u \in G$, and let us denote its degree, that is, the number of vertices adjacent to u, as d_u. The number of triangles passing through u is denoted by t_u. For each vertex u, we consider the pair (d_u, t_u).

When Algorithm 1 is applied to a particular graph G, given a vertex u, its *degree-triangles* pair (d_u, t_u) is transformed to another pair (d'_u, t'_u). Next proposition defines the *destiny region* of a pair (d_u, t_u), that is, the subset of $\mathbb{Z} \times \mathbb{Z}$ where the pair (d'_u, t'_u) may take its value.

Proposition 3.1. *Let G be a graph that has been perturbed into G' using Algorithm 1. Let (d_u, t_u) be the degree-triangles pair of u in G where $d_u \geq 1$, and let (d'_u, t'_u) be the corresponding pair in G'. Denoting by T and T' the number of triangles of G and G',* *respectively, the following inequalities hold:*

$$\max\{1, d_u - t_u\} \le d'_u \le d_u + T \ ;$$

$$\max\{0, d'_u - d_u\} \le t'_u \le \min\left\{T', \frac{d'_u(d'_u - 1)}{2}\right\}.$$

Proof. In the first round of Algorithm 1, edges of triangles are randomly removed until no triangles are left in G. For each edge removed from a triangle, the degree of its vertices is decreased by at most one. Hence, at the end of the first round, (d_u, t_u) is transformed to $(x, 0)$, where x is an integer that ranges between $\max\{1, d_u - t_u\}$ and d_u (the edge adjacent to a vertex of degree one will not be removed since it cannot be part of any triangle). Since the degree of a vertex cannot decrease during the second round, we obtain $\max\{1, d_u - t_u\} \le d'_u$.

After that, the second round of Algorithm 1 adds edges that create triangles. Each edge addition to the graph increases the degree of a vertex by at most one. Since the second round iterates at most T times, we get the degree of u cannot be more than $d_u + T$. At the end, the number of triangles passing through a vertex may be at most T'. Taking into account that a vertex of degree d cannot be part of more than $\frac{d(d-1)}{2}$ triangles, we get $t'_u \le \min\left\{T', \frac{d'_u(d'_u-1)}{2}\right\}$.

A vertex u with $d'_u > d_u$ implies that its degree has increased during the triangle addition phase. In this phase, each time u receives a new edge, the number of triangles passing through it also increases by at least one. This implies that $d'_u - d_u \le t'_u$. Since t'_u cannot take a negative value, we obtain $\max\{0, d'_u - d_u\} \le t'_u$.

□

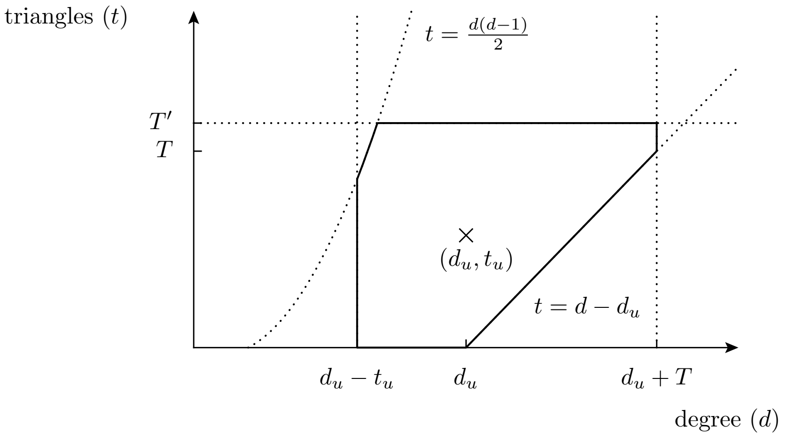

Figure 1: Destiny region D_u of (d_u, t_u) whose bounds are given in Proposition 3.1.

There are some special cases where the destiny region of a vertex can be better estimated. For instance, Algorithm 1 does not perturb isolated vertices ($d_u = 0$) nor connected components with just two vertices. Such simple structures are very common in social network graphs, and classical k-anonymity is directly provided on them when such isomorphic structures appear k times. In general, given a pair (d_u, t_u) from a graph containing T triangles, we can compute the bounds of its destiny region D_u. The size of this region depends on T and T'. Our experiments have shown that in practice $T \approx T'$ so that its size is $O(T^2)$. The shape of D_u is illustrated in Figure 1.

Not every point in D_u corresponds to a vertex u with (d_u, t_u), as illustrated by the following example.

Example:

Let $G = (V, E)$ be a graph with $V = \{a, b, c, d, e\}$ and $E = \{ab, ae, bc, be, cd, de\}$ (see Figure 2). G contains one triangle (a, b, e). Regarding vertex a, we have $(d_a, t_a) = (2, 1)$.

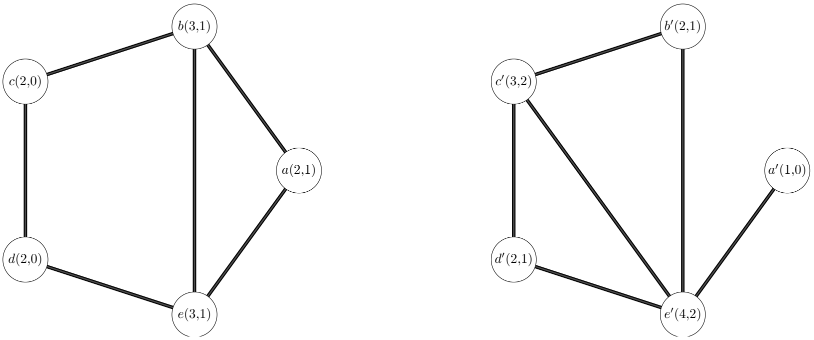

Figure 2: Graph G (left) and its anonymized graph G' (right) generated by Algorithm 1. In this case $T = 1$ and $T' = 2$. Numbers in brackets indicate the degree-triangles pairs.

From Proposition 3.1 we have $D_a = \{(1, 0), (2, 0), (2, 1), (3, 1), (3, 2)\}$. Algorithm 1 will first eliminate the triangle by randomly removing one of its edges. It will result in $(d_a, t_a) = (1, 0)$ with probability $\frac{2}{3}$ (either ab or ae are removed) and $(d_a, t_a) = (2, 0)$ (edge be is removed) with probability $\frac{1}{3}$. Next, one edge will be added so as to create a new triangle. It can be seen that at the end (d_a, t_a) becomes $(2, 1)$ with probability $\frac{2}{5}$; $(1, 0)$ with probability $\frac{1}{3}$; and $(3, 1)$ or $(2, 0)$ with probability $\frac{2}{15}$. Moreover, T' may become 2 in some cases (this is the case in G'), but the tuple $(3, 2) \in D_a$ is not possible for vertex a starting from its original value $(2, 1)$.

3.2 Measuring degree-triangle confusion

Let us consider an adversary whose goal is to re-identify a vertex u in an anonymized graph G' assuming her knowledge on u is given by the pair (d_u, t_u), that is, the adversary knows the number of relationships of u and the number of 'three party friends' that u belongs to.

Confusion is provided as long as the adversary has some level of uncertainty about the vertices of G' that may correspond to u. The set of candidate vertices is given by the set of vertices in G' that belong to the destiny region of u, namely D_u. The adversary does not know the value of T (number of triangles in the original graph G), but a good estimate is given by $T \approx T'$. If just one vertex of G' falls in D_u, then u will be re-identified resulting in the corresponding privacy compromise. The desirable situation is that in which the destiny region contains a large number of vertices, thus ensuring a high level of confusion. Let us define

$$\mathcal{M}_{G,G'}(u) = |\{v \in V \mid (d'_v, t'_v) \in D_u\}|,$$

as a local measure of confusion on u once G' has been released. Although not all vertices in D_u have the same probability of corresponding to u (as shown in the previous example), identifying u is not an easy task for an adversary with a limited knowledge of the original graph G. Hence, we propose the following confusion measure.

Definition 3.2 (Degree-triangle confusion)**.** Let $G = (V, E)$ and $G' = (V, E')$ be two graphs with a common set of vertices V. We say that (G, G') is a k-degree-triangle confusing pair of graphs if $\mathcal{M}_{G,G'}(u) \geq k$ for every $u \in V$. That is, every vertex in G has at least k matching candidate vertices in G'.

Note that the largest k satisfying $\mathcal{M}_{G,G'}(u) \geq k$ for each $u \in V$ corresponds to $\min_{u \in V}\{\mathcal{M}_{G,G'}(u)\}$. In the example of Figure 2, $\mathcal{M}_{G,G'}(b) = 4$ since vertices $\{b', c', d', e'\}$ fall in the destiny region of b (see Figure 3). For the rest of the vertices, we have $\mathcal{M}_{G,G'}(a) = 4$, $\mathcal{M}_{G,G'}(c) = \mathcal{M}_{G,G'}(d) = 3$ and $\mathcal{M}_{G,G'}(e) = 4$. Hence, (G, G') is 3-degree-triangle confusing.

Since $k = \min_{u \in V}\{\mathcal{M}_{G,G'}(u)\}$, it usually happens that the destiny region of most vertices contains more than k vertices. In order to provide more information about degree-triangle confusion, we will also analyze the median and the maximum values in $\{\mathcal{M}_{G,G'}(u) \mid u \in V\}$. The median value will provide the number of matching candidates for a "typical" vertex, meanwhile the maximum one corresponds to the worst case for an attacker.

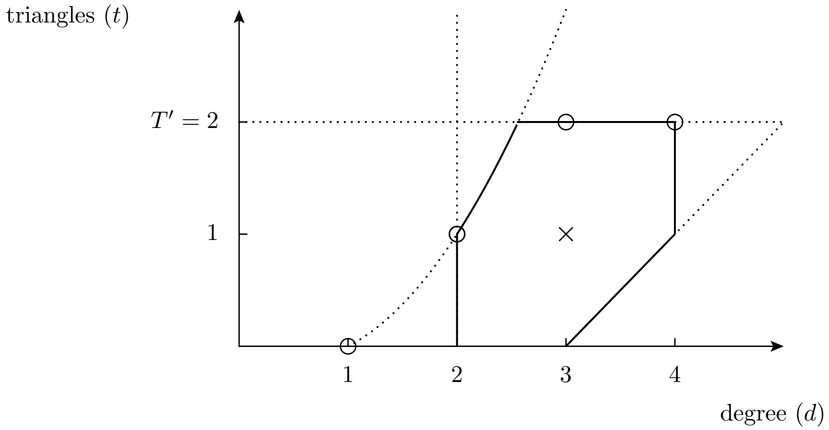

Figure 3: Destiny region of vertex b. Vertices of G' are represented by little circles (tuple $(2, 1)$ appears twice since it corresponds to vertices b' and d'). Vertex a', $(1, 0)$, is out of D_b.

4 Experimental results

The proposed anonymization procedure has been implemented in Python[1] using the Networkx[2] v1.7 graph library and tested over real and synthetic social network data.

4.1 Graph data sets

The following real data sets and synthetic graph generators have been employed in our experiments:

- The *Coauthors* graph is generated from bibliographic data available at the collection of Computer Science Bibliographies[3], where each author is assigned a vertex and edges are built from co-authorship relations. In our experiments, a graph with 11510 vertices and 11135 edges has been built. Data sets constructed in the same way have been previously used in [10, 20].
- The *Condmat* graph represents the collaboration network of scientists posting preprints on the condensed matter archive at `www.arxiv.org`. This version is based on preprints posted between January 1, 1995 and June 30, 2003. This graph is composed of 31163 vertices and 120029 edges. It has been used by several authors as a test-bed for community-finding algorithms for large networks (see, for example,[13]).
- The *Holme-Kim model* produces scale-free synthetic graphs in which the probability $P(k)$ that a vertex interacts with k other vertices follows a power law distribution, that is, $P(k) \sim C{\cdot}k^{-\gamma}$. Many real world graphs have a power law degree distribution with $2 \leq \gamma \leq 4$, as noted in [1]. The graphs generated by this model have $\gamma \approx 2.9$ which is considered a good approximation to many real world graphs. This model extends the well known *Barabasi-Albert* model (see [1]) including an extra step referred to as *triangle formation step*. In the *Barabasi-Albert* model, an empty graph with m vertices is first generated. After that, the construction algorithm iterates by adding a degree m vertex v at each step. Each edge of v is attached to an existing vertex with a probability proportional to its degree (this is the *preferential attachment (PA) step*). The Holme-Kim model incorporates an additional phase: for each edge between v and w added in the PA step, add one more edge from v to a randomly chosen neighbor of w with a given probability p. As a consequence, Holme-Kim graphs range between low-clustered graphs for $p = 0$ (Barabasi-Albert graphs) and highly-clustered ones for $p = 1$ (see [7]).
- The *Watts-Strogatz model* was inspired by the small-world phenomenon which is based on the notion that every person in the world is connected to anyone else through a chain of six mutual acquaintances at most (also known as “six degrees of separation”). Starting from a ring lattice on n vertices, where vertices have degree k, this model takes each edge and rewires it with probability p. This model interpolates between regularity ($p = 0$) and total disorder ($p = 1$). For $0 < p < 1$ we obtain highly clustered graphs having a small diameter, as it happens in many real world graphs (see [18]).

[1] `http://python.org`

[2] `http://networkx.lanl.gov`

[3] `http://liinwww.ira.uka.de/bibliography`

4.2 Graph utility measures

The method proposed in this paper introduces some changes to a graph prior to its release. In order to quantify the extent to which the original graph has been modified, we compute some statistical network measures and see how they are affected as a result of masking. These are the utility metrics.

The following utility measures have been considered:

- Number of links (*size*),
- Number of triangles,
- Average clustering coefficient,
- Average shortest path length (*Av. SPL*),
- Minimum, median and maximum vertex degree.

The (local) clustering coefficient c_u of a vertex u measures how close the neighbors of u are to being a clique, that is, $c_u = \frac{2t_u}{d_u(d_u-1)}$ when $d_u \geq 2$, and $c_u = 0$ otherwise. Given an order n graph $G = (V, E)$, the average clustering coefficient is,

$$\frac{1}{n} \sum_{u \in V} c_u.$$

Besides, the average shortest path length is,

$$\frac{1}{n(n-1)} \sum_{u,v \in V} \mathrm{dist}(u, v),$$

where $\mathrm{dist}(u, v)$ is the distance (length of a shortest path) between u and v. The other parameters are self-explanatory.

4.3 Experiments on real data sets

Algorithm 1 has been run ten times over the *Coauthors* and *Condmat* graphs. The median computation time has been 2.03 and 4.19 minutes, respectively.

Utility measures

Utility measures of the original graphs and their anonymized versions are shown in Table 1. As can be seen, the metrics of a graph and its perturbed versions exhibit a high correlation. After masking, the size is slightly increased while the median number of triangles is preserved although in some experiments it resulted in a slightly larger value (no more than four additional triangles were created in all the experiments). Degree parameters are well preserved in both graphs too. The average clustering and average SPL have been the most affected parameters in the *Condmat* graph. Figure 4 shows the distribution of (d_u, t_u) pairs for both graphs. Focusing our attention on the *Condmat* graph, it can be seen that vertices with high degree have less trianges after anonymization. As a consequence, the Av. clustering is reduced.

	Coauthors		*Condmat*	
	Original	Anonymized (median)	Original	Anonymized (median)
Size	11135	12040	120029	179244
Triangles	6395	6395	232994	232994
Av. Clustering	0.4591	0.4206	0.6488	0.47217
Av. SPL	3.3543	3.207	5.2995	4.2047
Min. Degree	0	0	0	0
Median Degree	2	2	25	27.5
Max. Degree	73	74	202	211

Table 1: Metrics of *Coauthors* and *Condmat* graphs before and after anonymization.

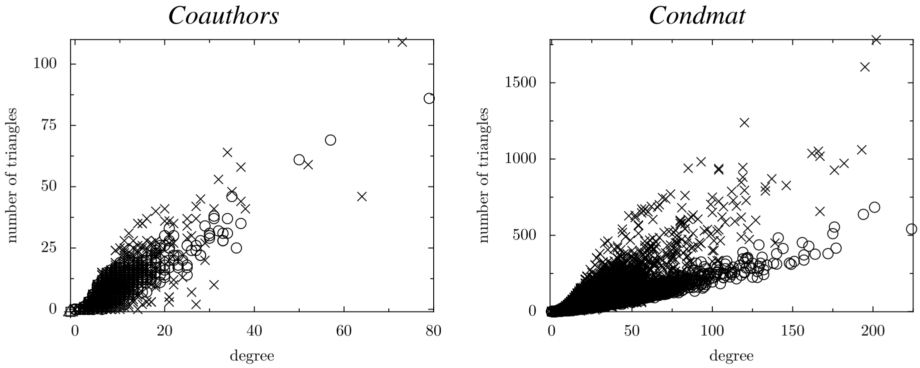

Figure 4: Degree-triangles distribution for *coauthors* and *condmat* graphs before (symbol $\times$) and after (symbol $\circ$) anonymization.

Degree-triangle privacy

The value of $\mathcal{M}_{G,G'}(u)$ has been measured individually for every vertex u after each experiment.

The *Coauthors* graph has 11510 vertices: 2574 of them are isolated vertices and 1982 are located in connected components with two vertices. These vertices are not affected by Algorithm 1 but they are indistinguishable in terms of re-identification. The remaining 6954 vertices are masked by Algorithm 1. Table 2 summarizes the results. As can be seen, the minimum value for $\mathcal{M}_{G,G'}(u)$ is 20, that is, (G, G') is a 20-degree-triangle confusing pair of graphs. Nevertheless, a 'typical' vertex has an elevated amount of vertices (6947) in its destiny region (more than 60% of vertices).

The *Condmat* graph contains 703 isolated vertices and 830 vertices belonging to connected components with two vertices. The remaining 29630 vertices have been masked by Algorithm 1. The destiny region of a 'typical' vertex contains 29630 vertices (95.1% of vertices). The smallest destiny region for any node includes 7198 vertices, that is, 24.3% of vertices (see Table 2). In this example, the degree-triangle confusion parameter is $k = 703$ which comes from isolated vertices.

	Values in $\{\mathcal{M}_{G,G'}(u) \mid u \in V\}$					
vertices u such that (d_u, t_u) satisfies	*Coauthors* (11510 vertices)			*Condmat* (31163 vertices)		
	Min.	Med.	Max.	Min.	Med.	Max.
$d_u = 0$	2574	2574	2574	703	703	703
$d_u = 1$	1982	1982	1982	830	830	830
$d_u > 1$ or $t_u > 1$	20	6947	6954	7198	29630	29630

Table 2: Minimum, median and maximum values of $\{\mathcal{M}_{G,G'}(u) \mid u \in V\}$ for *Coauthors* and *Condmat* graphs. The degree-triangle confusion measure for each graph is the minimum value in the table.

4.4 Experiments over synthetic graphs

Experiments on synthetic graphs have been performed over graphs with 10^5 vertices. For each model, we have generated ten random graphs with parameter p taking the following values: $0, 0.2, 0.4, 0.6, 0.8, 1$. Regarding Watts-Strogatz graphs, they have been generated from cubic graphs so that the resulting graphs have a minimum degree equal to three, except for $p = 0$, where the resulting graph is regular of degree 6. In all the cases, the generator provided by the Networkx library has been employed. Each graph from the test set has been masked ten times and the utility and privacy metrics have been computed (average values are analyzed).

Utility measures

The computation time of Algorithm 1 on both models is depicted in the left graphic of Figure 5. It can be seen that the computation time is strongly correlated with the number of triangles of the graph (right graphic of Figure 5). For a Watts-Strogatz (WS) graph of order 10^5 containing 300000 triangles, Algorithm 1 takes a little bit more than 11 hours. When $p = 0.5$, both models generate graphs with a close number of triangles (50000) and a similar computation time (around 4 hours) is required for both models. The number of triangles in the anonymized graphs almost equals the amount of triangles of the original ones, as it can be seen from the overlapping lines in the right graphic of Figure 5. More specifically, the maximum difference of $T' - T$ has been 6, which is a negligible quantity for graphs containing 300000 triangles.

The number of edges of the generated Watts-Strogatz (WS) graphs has been $3 \cdot 10^5$ in all the cases. The amount of edges after anonymization has been increased in all the cases, except for $p = 0$, where the anonymized graph contains 299591 edges (a 0.001% difference). The maximum difference appears for $p = 0.2$, where the corresponding anonymized graph has 322052 edges. Nevertheless, the maximum difference is about 7.35% of edges. Regarding Holme-Kim (HK) graphs, their size is close to $2 \cdot 10^5$. The HK anonymized graphs contain more edges than the corresponding value for the original ones, but again the maximum relative difference (for $p = 0.6$) is 2.07% (See Figure 6).

Degree parameters variation appears in Figure 7. It can be seen that the maximum difference of WS graphs appears for $p = 0.2$, where the maximum degree has been doubled (from 12 to 24), the median degree increases from 5.5 to 7.5 and the minimum degree

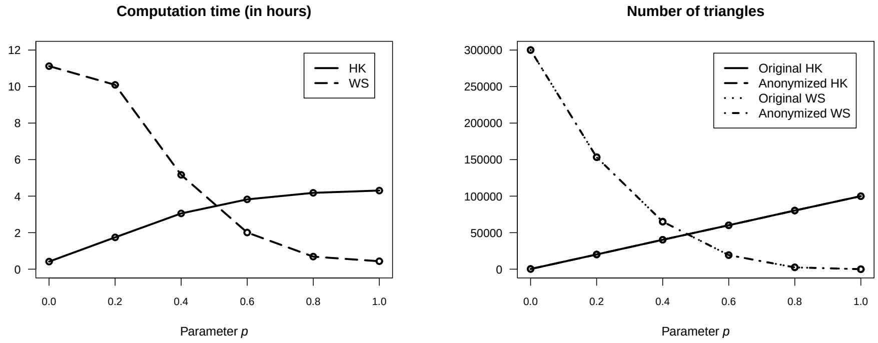

Figure 5: Computation time (left) and number of triangles for Watts-Strogatz (WS) and Holme-Kim (HK) graphs with 10^5 vertices, as a function of parameter p, before and after anonymization (right).

decreases from 3 to 1. Regarding HK graphs, degree parameters have been less affected.

The clustering coefficients in both models have been preserved to a high extent (see Figure 8). In WS graphs, the difference tends to zero as $p \to 1$. Besides, we observe the reverse behaviour in the HK model where the maximum difference appears at $p = 1$ where anonymized HK graphs are less clustered.

Degree-Triangle privacy

The minimum, median and maximum values of the degree-triangle confusion measure have been computed in both models. The results are shown in Figure 9 and Table 3. For Holme-Kim graphs, the privacy has been compromised for $p = 0, 0.4$ and 0.6, where $\min\{\mathcal{M}_{G,G'}(u) \mid u \in V\} = 1$, that is, there is at least one re-identifiable vertex in these cases. Holme-Kim graphs have a few vertices with high degree containing a small number of triangles. These vertices are difficult to mask, specially when the number of triangles T of the whole graph is also low. Masking data sets with outliers is known to be a thorny issue [11]. Prior to releasing a masked graph containing such nodes, some additional measures such as removing them or adding some additional noise should be taken. Nevertheless, a 'typical' (median) vertex contains an acceptable number of vertices in its destiny region, as Table 3 shows.

Regarding the Watts-Strogatz graphs, we have $\min\{\mathcal{M}_{G,G'}(u) \mid u \in V\} = 1$ just for $p = 1$. WS graphs have a low number of triangles (only 20 in our experiment) for $p = 1$. As a consequence, Algorithm 1 produces an insignificant modification to WS graphs so that vertices with a high degree are probably not affected and become easily re-identificable in the masked graph. Nevertheless, a 'typical' vertex contains a high number of vertices into its destiny region (7276), as Table 3 and the dashed line in left graphic of Figure 9 shows. As could be expected, these examples show the presented algorithm is not adequate for graphs having a reduced amount of triangles.

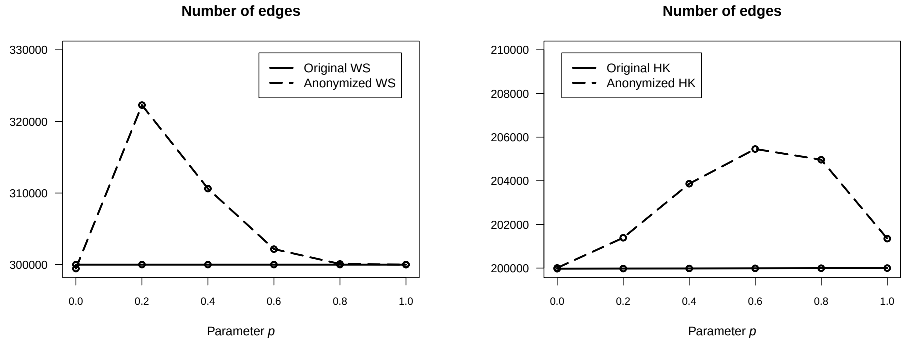

Figure 6: Number of edges for Watts-Strogatz (WS) (left) and Holme-Kim (HK) (right) graphs with 10^5 vertices, as a function of parameter p, before and after anonymization.

	Values in $\{\mathcal{M}_{G,G'}(u) \mid u \in V\}$					
Parameter p	Watts-Strogatz			Holme-Kim		
	Min.	Med.	Max.	Min.	Med.	Max.
0	10^5	10^5	10^5	1	13	70087
0.2	12368	90486	10^5	2	320	91885
0.4	271	59702	99990	1	968	95642
0.6	43	36453	98897	1	2278	99163
0.8	6	19392	87726	23	6489	99906
1.0	1	7276	44654	86101	99999	10^5

Table 3: Minimum, median and maximum values in $\{\mathcal{M}_{G,G'}(u) \mid u \in V\}$ for *Watts-Strogatz* (WS) and *Holme-Kim* (HK) graphs.

5 Conclusion and future work

In this paper a new anonymization method for social network graph data has been presented. The method is composed of two differentiated phases. The first phase iterates by randomly removing one edge of a randomly selected triangle in the graph until no triangles are left. In the second phase, the removed triangles are randomly reallocated in the graph. Due to its simplicity and low cost of the required operations, the method can be efficiently implemented in an algorithm whose running time grows linearly with the amount of triangles.

Empirical experiments have shown the method provides a high privacy level. Regarding data quality, experiments have shown structural parameters are better preserved in graphs with a larger homogeneity among vertices.

Some open issues that will be addressed in future research are:

- Quantify the probability distribution of the degree-triangles pair in the destiny region.
- Find techniques to increase structural parameters preservation in non-homogeneous graphs.

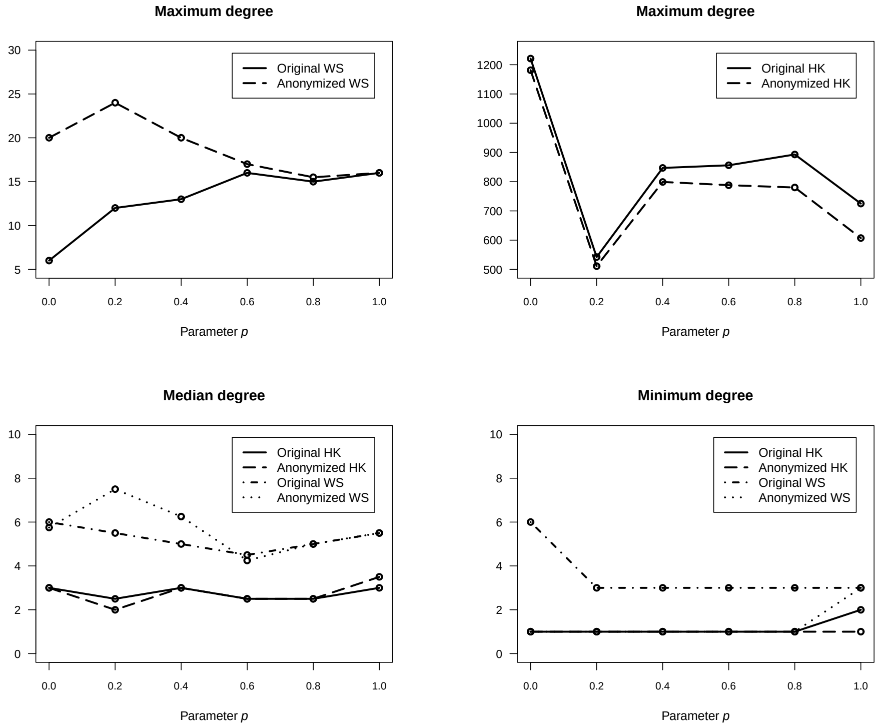

Figure 7: Maximum, median and minimum degree for Watts-Strogatz (WS) and Holme-Kim (HK) graphs with 10^5 vertices, as a function of parameter p, before and after anonymization.

Acknowledgements

The authors thank the referees for their constructive and helpful comments, which led to several improvements to the manuscript.

References

[1] A. Barabási and R. Albert, Emergence of scaling in random networks, *Science* **286** (1999) 509–512.

[2] A. Campan and T. M. Truta, A clustering approach for data and structural anonymity in social networks, *Proc. of PinKDD'08*, 2008.

[3] J. Cheng, A.W-C. Fu and J. Liu, K-Isomorphism: privacy preserving network publication against structural attacks, *Proc. of SIGMOD'10*, 2010.

[4] N. A. Christakis and J. H. Fowler, The spread of obesity in a large social network over 32 Years, *The New England Journal of Medicine*, **357**, (2007), 370–379.

[5] J. H. Fowler and N. A. Christakis, Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study, *British Medical Journal*, **337**, No. a2338, (2008), 1-9.

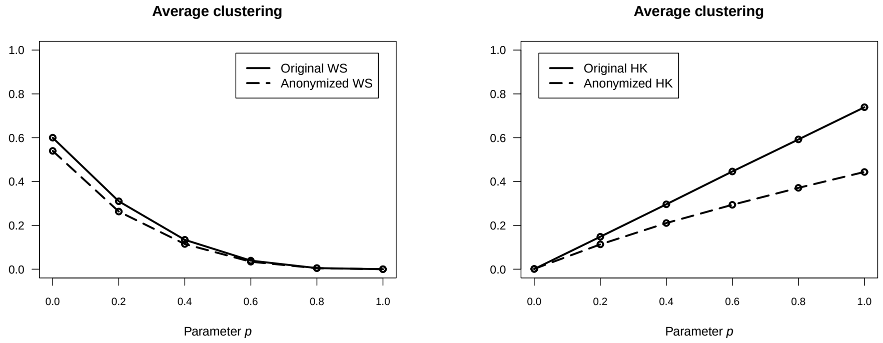

Figure 8: Average clustering for Watts-Strogatz (WS) and Holme-Kim (HK) graphs with 10^5 vertices, as a function of parameter p, before and after anonymization.

[6] M. Godwin, Maurice victim target networks as solvability factors in serial murder, *Social Behavior and Personality: an international journal*, **26**, No. 1, (1998) , 7583.

[7] P. Holme and B. J. Kim, Growing scale-free networks with tunable clustering, *Physical Review E.*, **65**, (2002)

[8] M. Hay, G. Miklau, D. Jensen, D. Towsley and P. Weis, Resisting structural re-identification in anonymized social networks, *Proc. of VLDB'08*, 2008.

[9] R. Kumar, J. Novak and A. Tomkins, Structure and evolution of online social networks, *Proc. of KDD'06*, 2006.

[10] K. Liu and E. Terzi, Towards identity anonymization on graphs, *Proc. of SIGMOD'08*, 2008.

[11] J.M. Mateo-Sanz, F. Sebé and J. Domingo-Ferrer, Outlier protection in continuous microdata masking, Proc. of PSD'04, Lecture Notes in Computer Science, **3050**, 2004, 201–215.

[12] L. Mercken, T. A. B. Snijders, C. Steglich, E. Vertiainen and H. De Vries, Smoking-based selection and influence in gender-segregated friendship networks: a social network analysis of adolescent smoking, *Addiction*, **105**, Issue 7, (2010), 12801289.

[13] M. E. J. Newman, The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci. USA 98*, (2001), 404–409

[14] P. Samarati and L. Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and supression, Tech. rep., SRI Intl. Tech. Rep., 1998.

[15] J. Scott, *Social Network Analysis Handbook* (2nd ed.), Sage Publications Ltd., 2000.

[16] K. Stokes and V. Torra, n-Confusion: a generalization of k-anonymity, Proc. of 5th Intl. Workshop on Privacy and Anonymity in the Information Society, 2012.

[17] S. Wasserman and K. Faust, *Social Network Analysis: Methods and applications (Structural Analysis in the Social Sciences).* Cambridge University Press, 1994.

[18] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, *Nature*, **393** (1998), 440–442.

[19] W. Wu, Y. Xiao, W. Wang, Z. He and Z. Wang, K-Symmetry model for identity anonymization in social networks, *Proc. of EDBT'2010*, 2010.

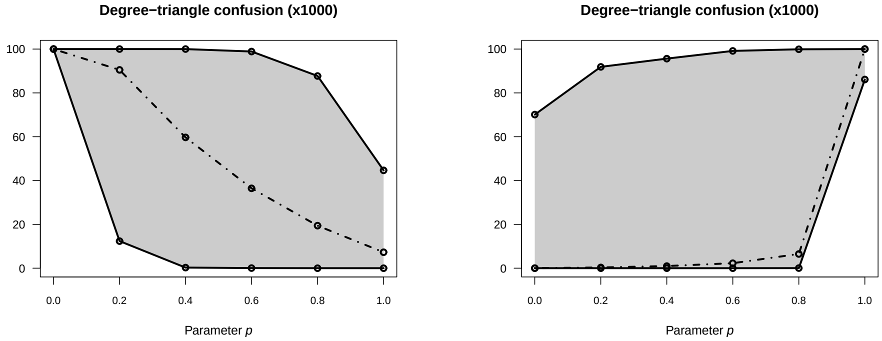

Figure 9: Degree-triangle confusion measure for Watts-Strogatz (left) and Holme-Kim (right) graphs with 10^5 vertices, as a function of parameter p. The dashed line indicates the value of $\mathcal{M}_{G,G'}(u)$ for a typical vertex, while the borders of the dark region are the maximum and the minimum values of the confusion measure $\mathcal{M}_{G,G'}(u)$.

[20] L. Zou, L. Chen and M. T. Öszu, K-Automorphism: a general framework for privacy preserving network publication, *Proc. of VLDB'09*, 2009.

[21] B. Zhou and J. Pei, Preserving privacy in social networks against neighborhood attacks, *Proc. of ICDE'08*, 2008, 506–515.