Measuring Local Topological Anonymity in Social Networks

Gábor György Gulyás
Department of Telecommunications
Budapest University of Technology and Economics
Budapest, Hungary
gulyasg@hit.bme.hu

Sándor Imre
Department of Telecommunications
Budapest University of Technology and Economics
Budapest, Hungary
imre@hit.bme.hu

***Abstract*—Service providers of social network based services release their sanitized graph structure for third parties (e.g., business partners) from time to time. However, as these releases contain valuable information additionally to what is publicly available in the network, these may be targeted by re-identification attacks, i.e., where an attacker tries to recover the identities of the nodes that were removed during the sanitization process. One powerful type of these, called structural re-identification attacks consider only structural properties, and work according to a specific strategy: first they re-identify some nodes by their globally unique properties, and then in an optional second phase, nodes related to these are re-identified by their locally unique properties. Global re-identifiability or global node anonymity is a well studied concept, however, node anonymity for local re-identification has not yet been analyzed.**

Therefore in this paper, after discussing the related literature on anonymity and re-identification, we introduce the novel term of Local Topological Anonymity (LTA), which describes the resistant power of a node against local re-identification attacks, or, in other words, indicates how well the node is structurally hidden in her neighborhood. Regarding these attacks in the literature, we propose three measure variants of LTA based on structural similarity measures, and evaluate them by visual inspection and simulation in multiple networks. We show that one of the proposed measures provides good prediction on local node re-identifiability as there is correlation between the LTA values and the re-identification statistics provided by the state-of-the-art algorithm.

Keywords- social networks, anonymity, re-identification

I. INTRODUCTION

Social network-based services play an essential part in the everyday life of many. One common feature of these services is that they all have an underlying graph structure, where nodes and edges can have different interpretations depending on the type of the service. For instance, in case of an online social networking service, a user can create a profile for herself (i.e., a node), and mark other users as her acquaintances (i.e., creating edges). Some services do not reveal directly the graph structure beneath them, though they should be considered social networks similarly, as in the case of mobile phone use or e-mail correspondence.

These services may be valuable sources of information for several related parties, such as application developers, researchers, or business partners. Therefore, if the graph structure of such a service is released in anonymized form but with additional private information, it might be in the interest of a related party (who is now considered as an attacker since she violates user privacy) to re-identify users in the anonymized data set. Additionally, users may create instances in numerous services in parallel, and the attacker may try to link related instances. In both cases, the attacker builds the de-anonymization attack on his a priori knowledge (or auxiliary information), which can be some properties of the targeted users or even partial crawls of networks.

However, de-anonymization is not a trivial task, as user instances could be published with differing or contradictory profile information under unrelated identifiers. In order to avoid false identification, it has been shown that structural properties can be used as alternative source of correlation [1-6], which methods are called structural re-identification or de-anonymization attacks. By considering the extent of graph structure modification prior to sanitization, re-identification attacks can be categorized as active [1] or passive [2-6]. In active attacks, the malicious party modifies the network structure to find the specific injected subgraph (and the nodes it is connected to), while passive attacks use auxiliary information solely. For structural passive re-identification attacks, auxiliary information can be degree values of the node's neighborhood [3], the structure of the neighborhood [4], or another graph that overlaps with the sanitized one [5, 6].

Concerning the strategy of the attack, re-identification can be carried out in two sequential phases: the global and the local re-identification phases (also called seed identification and propagation, respectively in [5, 6]). In the first phase, nodes (or subgraphs) with globally unique structural properties are re-identified, and after the appropriate number of nodes are de-anonymized, the second phase can be started. In this phase, nodes connected to

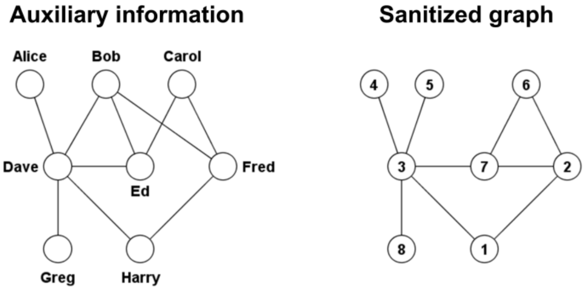

Figure 1. The perturbed, anonymized graph to be de-anonymized (right) and the attacker knowledge crawled from another service (left).

already re-identified nodes are being analyzed, and the ones with locally outstanding structural properties are re-identified.

For example, an attacker tries to de-anonymize users based on their degree values in a sanitized graph, assuming her inputs are as it is depicted on Fig. 1. Regarding global re-identification, in the sanitized graph nodes form anonymity sets {4, 5, 8}, {6, 1}, {7, 2}, {3}, so the attacker who is looking for Harry, cannot re-identify him directly from {6, 1}. Therefore, she starts with global re-identification, and de-anonymizes node 3 as Dave. Then she proceeds with local re-identification, as she knows that Harry is connected to Dave and his degree value is 2, therefore she re-identifies node 1 as Harry, from the candidate set of {1, 4, 5, 7, 8}, i.e., local nodes around re-identified node 3.

Most structural de-anonymization techniques only implement the global re-identification phase [1-4], limiting their success rate to small scales, as it is not computationally feasible to de-anonymize hundreds of thousands of nodes this way. However, it has been shown that adding the second phase opens the way to efficient large-scale re-identification [5, 6]. Similarly as in the previous example, these attacks use two overlapping graphs as their input, a sanitized graph as their target (i.e., the target graph), and an auxiliary graph containing identifying information (i.e., the source graph), and they flag a target node as re-identified if it is assigned to a source node.

In this paper, our focus is on the second, local re-identification phase, and as our main contribution, we introduce a novel anonymity measure called Local Topological Anonymity (LTA) for this phase, expressing the node's resistance level against local re-identification techniques. The LTA measure is beneficial for users to apply privacy-enhancing software to establish their privacy status more accurately than simply considering their global structural re-identifiability (which app can also suggest making modifications in their relationships to achieve stronger privacy); by measuring their level of discoverability against the state-of-the-art structural re-identification attacks, i.e., ones implementing a local re-identification phase. In addition, the LTA gives a fast to calculate a posteriori anonymity measure, and it requires inputs that can be reasonably assumed to be available for users in most services: the neighbors and neighbors of neighbors of nodes.

Besides, the LTA measure also allows data providers and attackers to estimate the possible success of attacks; e.g., in deciding whether a dataset (a network graph) is ready for release or needs modification (e.g., low median LTA value), or worth attacking. In case of particular nodes that cannot be identified globally, LTA values also offer the possibility to check prior the release or the attack. Additionally, an attacker can calculate the LTA values of multiple users to locate anonymity and similarity sets, and check whether certain users are part of them or not. In this paper we focus on the anonymity measure of nodes, and leave the detailed analysis of measures calculating an LTA value of the entire network for future work.

The paper is structured as follows. In Section II, we discuss the related literature of structural anonymity and re-identification. In Section III, we introduce the concept of LTA, discuss the related theoretical foundations, and propose three variants of LTA measures. We present the evaluation of the measure afterwards in Section IV, first by visually comparing the results of the measures given in small networks, and then by simulation in larger networks. Finally, in Section V, we conclude our work and discuss issues for future work.

II. Anonymity in Social Networks

In social networks, anonymity can be characterized both for the whole network, and for a single node (or for a set of nodes). Our work focuses on node anonymity, but the work of Sing and Zhan is a good example for the prior: they define a graph structural measure, called topological anonymity [7]. Their measure describes variance of patterns in the complete graph with a single, normalized real number, based on node degree anonymity sets: the output value incorporates the variance of the clustering coefficients of the nodes in the sufficiently large sets. Instead of identity disclosure, their work concerns link disclosure, i.e., when an attacker is interested in the presence or absence of an edge between certain nodes.

Related to node anonymity, but before the research of networks and re-identification attacks, the term structural equivalence appeared in sociology [8]. Although based on similar principles, structural equivalence is too rigid within the current context: for structural re-identification two nodes having the same in- and out degree values can be globally undistinguishable, while considered structurally equivalent if they connect to the same neighbors with their incoming and outgoing edges [8]. Additionally to structural equivalence, nodes can be compared structurally in a more sophisticated way with similarity measures. These measures originate from other areas of science, but are also used in social networks for structural comparison of nodes in recommendation systems [9] and in re-identification techniques, e.g., cosine similarity in the second phases of the algorithms in [5, 6].

Practical structural node anonymity measures are reflecting the node's hiding ability against certain re-identification schemes, and are defined accordingly. Most of such measures are based on structural uniqueness or similarity measures, e.g., a node is considered anonymous considering its surrounding subgraph if there are a number of equivalent or similar coexisting facsimiles present in the graph, where their number may be a limit to count these similar structures to be an anonymity set [2-4].

Several variants of global re-identification techniques are derived from the concept of k-anonymity [10]. Zhou and Pei define that a node is k-anonymous if there are at least (k-1) other nodes with a similar neighborhood [4] (including adjacent nodes only). Liu and Terzi define k-degree anonymity similarly: a node is k-degree anonymous, if there are at least (k-1) other nodes with the same degree values in the network [2]. Hey et al. define k-candidate anonymity that relates anonymity of a node to the number of its mapping candidates [3], and give the analysis of three different types of global re-identification attacks: vertex refinement queries (a node is identified by the degree values of its neighbors, or neighbors of neighbors), subgraph queries (a node is identified by a surrounding subgraph described by the implied edges), hub fingerprint queries (a node is identified by its relation to specified hub nodes).

Structural node anonymity is not explicitly measured in all related works. Backstrom et al. present global re-identification phases (an active and a semi-active) that attempt to form a unique subgraph in the graph prior to anonymization [1]. In their work the structural uniquness (and identifiability) of the injected subgraph is controlled by the malicious third party, whose goal is to create a unique, but not trivially outstanding structure. Narayanan and Shmatikov use a uniqueness criterion for 4-cliques in the first phase of their passive algorithm in [5], and cliques being similar enough are considered to be forming an anonymity set (although clique uniqueness is only considered).

Above mentioned identification techniques consider only global structural uniqueness during the re-identification process. As a result, these techniques are only capable to re-identify the nodes in the structure that the algorithm is looking for (along with the connected nodes), and therefore lack the possibility of large-scale re-identification. As a solution, a local re-identification algorithm was introduced by Narayanan and Shmatikov [5]: after de-anonymizing some nodes by their globally unique structural properties, nodes in their neighborhood are de-anonymized by their structural properties that make them locally outstanding and unique. In their work they de-anonymized successfully 30.8% of the co-existing nodes in the graphs crawled from of two different services (Flickr and Twitter). In a second experiment where both graphs were obtained from the same service (but at different times), an even higher re-identification rate was achieved with a different global and a modified (but conceptually similar) local re-identification phase [6]. In both cases, the local re-identification phase considers only the 2-neighborhood of a node while trying to re-identify new nodes.

III. Local Topological Anonymity

To the best of our knowledge no work in the literature has yet considered measuring structural anonymity regarding the local re-identification phase. Therefore we introduce *Local Topological Anonymity* (*LTA*), which represents the level of resistance of a node against attacks considering local structural information only, or, in other words, describes how efficiently a node is hidden in its neighborhood regarding her structural properties locally.

In order to describe LTA measure in more details, we take a look into the propagation phases of attacks in [5, 6] to clarify their (common) basic principles. Both phases are round based, and in each round the algorithm starts with a set of already re-identified nodes, and tries to extend this set by de-anonymizing new nodes. In each round the algorithm iterates through all source nodes: a source node's re-identified neighbors assigned pairs (in the target graph) are selected, and their neighbors are the possible candidates of the source node. Then, the source node and the candidate nodes are compared by their structural properties, and a score is assigned for each target candidate. If there is an outstanding candidate score, the corresponding target node is assigned to the source node, and considered to be re-identified. Additionally, as re-identification spreads across the network, already identified nodes are revisited to perform corrections.

The comparison is the most fundamental part of these algorithms: a source node can be re-identified successfully if it is present in the target graph, and the local structural properties of the source node and the corresponding target node are similar enough, and its score makes the target node more similar than others in its neighborhood (i.e., outstandingly unique). In our opinion, possible propagation phase algorithms appearing in the future will share these principles too (and may involve additional ones), and therefore we base the definition of LTA measures on the structural uniqueness of a node in her neighborhood.

Definition 1. A Local Topological Anonymity measure is a function, denoted as $LTA(\cdot)$, which represents the hiding ability of a node in a social network graph against attacks considering only the structural properties of the node, within its d-neighborhood[1].

In this section, we discuss the theoretical basis of LTA measures, and propose three measure variants, which are evaluated in the following section.

A. Theoretical Basis and Initial Tests

The hiding ability of a node can also be interpreted based on its structural similarity to others in her neighborhood. In other words, if there are structurally similar nodes in its neighborhood, a node has a better chance for avoiding re-identification. Furthermore, the more similar nodes there are, the higher are the chances. (Here, we consider undirected networks for the sake of simplicity, but the concept can be extended for directed networks also – we leave this issue as future work.)

Therefore, as the basis of LTA measures, we propose similarity measures applied to the structural context. N.b. the similarity measure should be tailored for the structural property considered by the attacker, such as in the case of the example provided at the end of this section. For regular similarity measures (e.g., used for calculating the similarity of sets), there are several choose from, but the following rational considerations can help to filter out many of these:

- Fast to calculate. Calculating a single LTA value may involve numerous similarity measurements (e.g., calculated for all neighbors of neighbors), making LTA calculation a costly operation in case of slow measures.
- Not recursive. For similar reasons as above, recursive similarity measures can be costly too, and in addition, the complete graph may not be known at all times.

[1] We consider LTA measures of 2-neighborhoods only, as $d \geq 3$ may be impractical because of typically small network diameters.

- Positive values. Results derived from positive values can be interpreted and compared easier (e.g., when summing values).
- Normalized values. Normalized values of different nodes can be compared easier, and in addition, the average of normalized values is also normalized.
- Symmetric. The similarity value should not change when the order of the nodes is altered.

Respecting these considerations, there are still multiple similarity measures to choose from; however, the cosine similarity (which fulfills them all) outstands for two reasons. First, in their research, Spertus et al. compared six distinct similarity measures on the Orkut social network database, and found cosine similarity to give the best results [9]. In their experiment, similarity of users was calculated based on their community subscriptions regarded as sets, which is quite similar to our current case, where a node (a user) can be regarded as a set of its neighbors, and two nodes can be compared accordingly. On the other hand, both in propagation phases of [5, 6] the comparison mechanisms are based on cosine similarity: the scores are derived from the cosine similarity of the source node and each of the target candidates. Therefore we choose cosine similarity, which is denoted in set notation as

$$CosSim(v_i, v_j) = \frac{|V_i \cap V_j|}{\sqrt{|V_i| \cdot |V_j|}},$$

where v_i, v_j are nodes in the graph, and V_i, V_j are the sets of their neighbors respectively.

However, at the beginning of our empirical experiments, because of curiosity and for the sake of completeness, we compared some other measures to cosine similarity. Measures were tested with the LTA_B measure variant (see Section III.B for details), and some produced outputs relatively close to cosine similarity. Here we mention some of these alternatives; although this is not a closed list, and there might be several more alternative similarities to be chosen from.

While Spertus et al. consider the Pearson similarity measure [8] to be inappropriate for the current use [9], and in addition it produced negative LTA values in our experiments, it seems to be a good alternative in larger networks (starting from around thousand nodes): the correlation with cosine similarity was almost perfect for all larger test networks, namely 99.8% and above (for the data sets see Section IV.A).

For all test networks in all sizes, the Jaccard similarity [11] provided very close results to cosine similarity. Though some other measures also gave good results, their correlation degraded as the network sizes grew. This was the case for the L1-Norm similarity measure [9], the pointwise mutual-information similarity measures [9], and the similarity measure with minimum normalization (called topological overlap in [12]). Interestingly, and against our expectations, the asymmetric Salton IDF similarity measure [13] had convincing results also like the latter ones.

B. Proposals for LTA Measures

In the proposed LTA measures, the similarity of a node to its 2-neighborhood is based on the similarities to her neighbors of neighbors (in case of triangles, both neighbor nodes are considered as neighbor of neighbors, too). As a basis, the sum of similarities between v_i and all neighbors of neighbors is calculated, but normalized differently for each variants.

The first (and the simplest) measure variant, denoted as $LTA_A(v_i)$, is simply calculated as the average similarity of node v_i and its neighbors of neighbors. While constructing $LTA_B(v_i)$ the goal was to penalize nodes with high degree values, thus the sum of similarities is normalized with the degree of v_i. To cut back LTA values for one-degree nodes, the minimum value for normalization is 2 (during the visual comparison, it produced better results). The third measure, $LTA_C(v_i)$ is based on $LTA_A(v_i)$, but it penalizes nodes where node degrees in the set of the neighbors of neighbors vary greatly.

Therefore, we define these variants of LTA measures as follows:

$$LTA_A(v_i) = \sum_{\forall v_k \in V_i^2} \frac{sim(v_i, v_k)}{|V_i^2|}, \qquad (1)$$

$$LTA_B(v_i) = \sum_{\forall v_k \in V_i^2} \frac{sim(v_i, v_k)}{max(|V_i|, 2)}, \qquad (2)$$

$$LTA_C(v_i) = \sum_{\forall v_k \in V_i^2} \frac{sim(v_i, v_k)}{|V_i^2| \cdot max(\sigma_{deg}(\Delta V_i^2), 1)}, \qquad (3)$$

where V_i denotes the set of neighbors, V_i^2 denotes the set of neighbors of neighbors for node v_i, $sim(v_i, v_k)$ stands for the similarity function between two nodes (cosine similarity in our experiments), and $\sigma_{deg}(\Delta V_i^2)$ denotes the standard deviation of delta in degree values between v_i and members of V_i^2, i.e., $\Delta V_i^2 = \{\forall v_k \in V_i^2 : ||V_i| - |V_k||\}$.

We remark that a node may be well hidden according to a given LTA measure, but it may be still vulnerable to global re-identification attacks. This should be considered in some cases, like for small networks, where brute-force global re-identification attacks are feasible, and also for nodes connecting to unique subgraph structures formed by other participants (e.g., multiple hubs). Consequently, an LTA value assigned to a node should not be interpreted as a comprehensive overview of its anonymity status, but rather a complementary indicator regarding specific local re-identification phase attacks.

C. Example: Choosing a Similarity Measure and Calculating the LTA value

Here, we give an example for calculating LTA_A for the example (source) network given in the introductory section, depicted on Fig. 1; although this network is great for a simple example, we emphasize that LTA measures are useful complementary indicators for larger networks, where global re-identification is not feasible.

In the given example, the attacker uses node degree values for comparison, therefore we propose calculating the difference between degree values as a similarity measure, as

$$DS(v_i, v_j) = \frac{\max(|V_i|, |V_j|) - \big||V_i| - |V_j|\big|}{\max(|V_i|, |V_j|)},$$

where values are normalized for the ease of comparison. Based on this measure, LTA_A can be calculated for Alice as follows:

$$LTA_A(v_A) =$$
$$= \frac{DS(v_A, v_B) + DS(v_A, v_E) + DS(v_A, v_G) + DS(v_A, v_H)}{4}$$
$$= \frac{0.33 + 0.33 + 1.0 + 0.5}{4} = 0.54,$$

and their local structural anonymity status can be calculated accordingly for other nodes, too.

IV. Analysis of the LTA variants

We evaluated the proposed LTA measures by checking the correlation between LTA values and re-identification rates by the simulation of the propagation phase. Although the propagation phase is meant for larger networks, we also checked the measures by visual inspection in small networks.

A. The Datasets

For the tests in larger networks, we used graphs that had at least a thousand nodes, as we found that size to be a rational compromise between data size (thus simulation runtime) and ones giving lifelike results for LTA measure distributions and simulation results. The main data sources[2] were the Epinions who-trust-who network (collected in 2003), and the Wikipedia vote network (collected until January of 2008), the Slashdot network (collected in February 2009). For the comparison of results, an additional LiveJournal graph containing approx. half million nodes were crawled at the end of 2010 (at our dept.). We used these networks to create test data for the simulations by exporting subgraphs of the desired sizes with breadth-first search, where the resulting graph sizes were optimized for simulation runtime (see Table I).

The final form of our test data were derived by applying the perturbation strategy proposed by Narayanan and Shmatikov [5] to these exports, which creates two graphs with the desired node and edge overlap factors (denoted respectively as α_V and α_E) without adding new nodes or edges, therefore results in realistic source and target graphs. We measured the strength of the overlap in the number of coexistent 4-cliques, as it limits the maximal strength of the simulated adversary: propagation phase simulations were initialized with random disjoint seeds chosen from those. We ran the perturbation algorithm on the six graphs to create all test datasets. Ten of the datasets were created from WV-1000 and EP-1000 with differing clique overlaps to analyze the success rate and other properties of the propagation phase for different overlap factors (see Table II).

TABLE I. Initial datasets and their parameters.

Network	**Nodes**	**Edges**	**Density**	**Diam.**	**Avg. path length**
EP-1000	1,000	29,509	0.0591	4	2.1746
WV-1000	910	9,407	0.0227	5	2.7708
SD-1000	1,104	10,348	0.0170	5	2.4295
LJ-1000	1,033	10,521	0.0197	4	2.5608
LJ-10K	10,056	231,416	0.0046	6	2.8291
WV-Full	7,115	100,762	0.0040	7	3.2475

TABLE II. Derived data sets for the experiments.

Variant	α_V	α_E	**Clique overlap**
WV-1000vA	0.5	0.60	117
WV-1000vB	0.5	0.75	469
WV-1000vC	0.5	0.90	1,403
WV-1000vD	0.4	0.75	115
WV-1000vE	0.6	0.75	1,264
EP-1000vA	0.4	0.45	1,536
EP-1000vB	0.4	0.60	2,093
EP-1000vC	0.4	0.75	12,271
EP-1000vD	0.5	0.60	10,495
EP-1000vE	0.6	0.60	27,227
SD-1000	0.5	0.85	1,844
LJ-1000	0.5	0.90	2,085
LJ-10K	0.4	0.60	2,422
WV-Full	0.5	0.6	4,699

B. Visual Comparison in Small Networks

We expected LTA measures to provide credible scores in small graphs, too (e.g., some manually constructed, and other small graphs available for download[3]). Thus we calculated the measures for small networks, and compared them visually with expectations such as:

- Anonymity sets. Nodes having exactly the same neighbors (i.e., structurally indistinguishable nodes) should get the same LTA values, that are significantly higher than the average in the.
- Similarity sets. Nodes having a relatively significant overlap (to their degree) in their neighborhood (i.e., structurally similar nodes) should get values close to each other, and values should be higher than the average.
- Nodes that are structurally unique in their neighborhood (e.g., local hubs) should get lower values than the average.

Measure LTA_B produced the most credible results in small networks in contrast to the results in larger networks (see details later). As displayed for the undirected version of Zachary's karate network (see top of Fig. 2), LTA_B highlighted multiple anonymity sets, from which the largest is denoted as A_1, and a similarity set is also visible, marked as S_1. As expected, LTA_B values are lower for structurally unique nodes, such as for peripheral nodes or hub nodes. Measures LTA_A and LTA_C most notably differ for hubs. For

[2] SNAP Large Networks: http://snap.stanford.edu/data/index.html

[3] Network data: http://www-personal.umich.edu/~mejn/netdata/

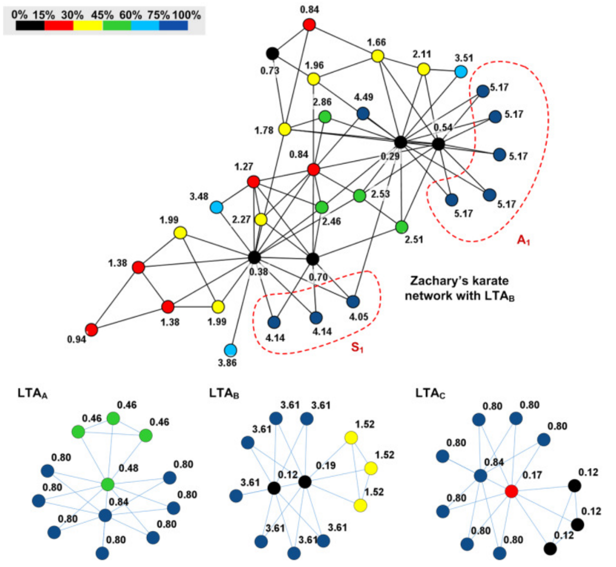

Figure 2. Examples for LTA values in small networks. Nodes are colored in accordance with their relative scores in the network, and absolute scores are displayed as numbers.

both measures, the node with the highest degree received the highest score in a small test network (see bottom of Fig. 2), higher than the largest anonymity sets. Furthermore, the two hub nodes are scored higher than nodes in the anonymity sets related to them. As the propagation phase is not designed for small networks, the simulation evaluation determines the real usefulness of the variants; however, all LTA variants gave appropriate scores for nodes in anonymity or similarity sets.

C. Choosing the Proper Algorithm and Settings for Simulation

Before the evaluation, we had to choose from the propagation phases in [5] and [6] for the simulations. First, there is a major difference in how the algorithms are used: the [5] was run on two graphs of different services, while the [6] was run on two snapshots of the same graph. Thus, we consider the [5] propagation phase to be more generic, and during simulations it seemed to be also more fault tolerant (which is not surprising according to the original use-cases). Although we could achieve similar results with both algorithms, the [5] propagation phase turned out to be easier to control (as it has less parameters), and parameters were less data dependent, too. As a result, we had chosen the propagation phase algorithm in [5] for our simulations.

In the simulations, we intend to measure the success rate for each node; however, we observed that the success rate depends heavily on further important settings, not just the algorithm itself. The authors mention that large-scale propagation depends on the seed size [5]; we confirmed their statement with simulations, and – in accordance with their results – we measured two thresholds for the seed set size (i.e., the number of 4-cliques): a lower where notable propagation really starts, and an upper after which additional seeds will not lead to significantly wider re-identification. Therefore, in the simulations we tried to use the largest number of seeds for initializing the algorithm, although the seed set size is limited by the number of overlapping cliques and networks structure, which is eventually assigned by the overlap factors of the nodes and edges.

As another interesting result, we found that the propagation phase is quite sensitive for seed locations. Not surprisingly, when a single 4-clique is used as a seed, the reachability horizon of the algorithm varies greatly for different seed locations, while the algorithm output is more or less deterministic. However, for an attacker using multiple seeds, seed location is still a problem: in our tests, it proved to be easier to achieve large-scale propagation with randomly selected seeds or with seeds constructed of low degree nodes; when all seed nodes had higher degree values, simulation achieved worse propagation results. Thus for initializing propagation, we used random seeds, selected from arbitrary parts of the network. (Although in reality, an attacker may only have high degree seeds, which is reckoned to be weaker than the simulated one leading to results similar in shape, but lower re-identification rates for most nodes.)

D. Simulation Details and Interpreting Results

As mentioned before, we evaluated the LTA measures by checking the correlation between the LTA values and re-identification statistics. Initially, we calculated LTA values for all tests network variants, and executed the [5] propagation phase multiple times. After trying various settings we found that 10 rounds of simulation gives a good average of re-identification rates (for these datasets).

In each round, a network dependent number of coexistent 4-cliques are selected randomly. To find a clique, first an arbitrary node is selected, and then a coexistent clique is chosen from its neighborhood. Afterwards, the simulation of the propagation phase initialized with the selected seeds. During the simulation, each node is assigned a re-identification score: if the node is re-identified successfully its score is increased by one (seeds are excluded from scoring), and in case of false re-identification it is decreased, otherwise the score is not changed. Then the re-identification statistics for the coexistent nodes are analyzed, and for comparison the results are plotted on XY graphs, and average of the re-identification statistics on line graphs (averages calculated for 50 LTA intervals), where the scores are ordered according to the given LTA measure.

For checking the correctness of each measure, we check the correlation between re-identification scores and LTA values. High (and negative) correlation can be observed if LTA values order re-identification scores in a decreasing order, or if the proportion of high and low re-identification values change in favor of low values, as LTA increases. However, it turned out that the prior is less frequent, as a node typically has a maximum or a minimum re-identification rate, not between (e.g., 83.9% of nodes fall in these cases for LJ-10K).

E. *Evaluation by Simulation*

In Fig. 4 we give two typical examples for the proposed LTA measures for EP-1000vC (9-10 disjoint seeds used for simulation) and WV-1000vC (7 seeds), and unordered results are also presented for comparison (i.e., re-identification rate ordered by node id). At first sight, correlation with LTA_A seems to be the best, other measures have some handicaps. For LTA_B, due to its heavy tailed distribution, there are significantly larger number of low LTA values, thus most of the results fall in the lower parts of the domain. Therefore LTA_B has more spikes for higher LTA values, since there are less nodes with low re-identifications scores, which could balance it in the average (see (1) on Fig. 4); this phenomenon occurred in most of the cases. Measure LTA_C produced apparently better results than LTA_B (but not if we consider correlation), even seemingly good correlation similarly as LTA_A, but not in all cases. As shown for EP-1000vC, it may start with low values (see (2) on Fig. 4), and the tail of the average function is not always descending (see (3) on Fig. 4).

For comparison, we also calculated the Pearson correlation coefficient for each test (see Fig. 3), and the average correlation values (for all networks) were $-0.421, -0.344, -0.269$, respectively for the proposed measures. The average of the highest achievable correlation was -0.814 (i.e., perfect ordering regarding the simulation results). Regarding these values, and the previously discussed characteristics, we had chosen LTA_A as the most appropriate measure (although LTA_B produced slightly better results in a few cases, it was less accurate in total), according to the results [5] propagation phase.

On Fig. 5 we give further examples of LTA_A and re-identification statistics for networks SD-1000vA (8 seeds), LJ-1000vA (11 seeds), and LJ-10KvA (30 seeds). As an interesting finding, it can be seen that it was not always possible to reach large-scale re-identification by simply using the highest number of available seeds; however, in other cases where it could be achieved, LTA values and re-identification statistics were more balanced, and the average function had a descending tail, similarly as seen on Fig. 4.

As displayed on the XY graphs and on the right-hand side of Fig. 5, re-identification scores significantly deviate from the average. This is for the reason we mentioned earlier: not that the re-identification scores are ordered in a descending manner, but for the proportion of positive and zero scores changing for each higher LTA intervals during the average calculation (in favor of zero scores; there is only a minority of negative scores). This is because if a node can be found by the algorithm, it is likely to be found in all simulation rounds, even for high LTA values. For instance, as in the re-identification statistics of LJ-10K, the number of nodes that were found zero and 10 times greatly outnumbered all other nodes (with proportions of 23.77% and 60.13% respectively). Furthermore, only the 1.29% of all coexisting nodes had negative re-identification scores, and only the remaining 14.81% of the nodes had a re-identification score between 0 and 10.

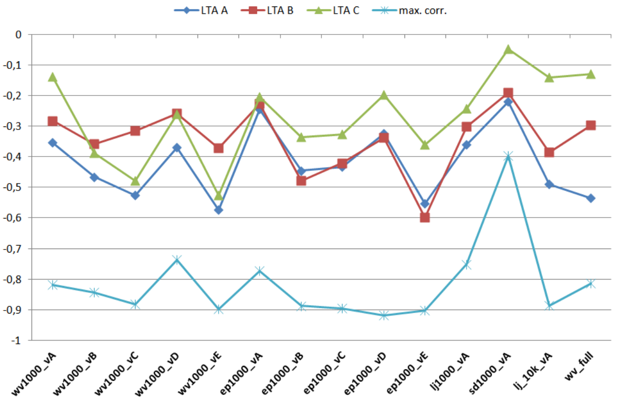

Figure 3. Pearson correlation coefficient values for different LTA measures in different networks.

V. Conclusion and Future Work

In this paper, after discussing the related literature and concepts of structural anonymity, we proposed a novel term, called Local Topological Anonymity describing the anonymity status of a node regarding local re-identification attacks. For measuring this aspect of anonymity, we proposed three LTA variants, and evaluated these measures with simulation, and as a result, we proposed one of the measures, namely the LTA_A, that allows the most accurate anonymity status estimation with respect to the state-of-the-art propagation phase algorithms. During the simulations, the LTA_A measure had the highest correlation with node re-identification statistics in average. Our novel measure is useful for the user measuring her anonymity status more accurately, and it may also be useful for an attacker or a data publisher to establish the possible success rate of an attack against certain nodes.

We mentioned some issues in the paper left for future work, such as creating a measure for providing estimation of the complete resistance level of a network, or to extend the currently proposed LTA measures for directed networks also. We also find the concept of a combined measure interesting, where the global and local re-identifiability of a node could be combined in a single indicator value (where one of the original components could be the LTA measure output). The combined measure could provide a comprehensive estimation describing the overall anonymity status of a node, for instance, regarding both a specified pair of seed identification and a propagation phase attacks. In addition regarding Fig. 3, there is also room for improving the currently analyzed measures, for instance by adding heuristics to the algorithm to achieve higher correlation.

Acknowledgment

Our work presented in this paper was supported from the KMOP-1.1.2-08/1-2008-0001 project by the BME-Infokom Innovátor Nonprofit Kft.

References

[1] Backstrom, L., Dwork, C., Kleinberg, J.: Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In: 16th international conference on World Wide Web, pp. 181-190, ACM, New York (2007)

[2] Liu, K., Terzi, E.: Towards identity anonymization on graphs. In: ACM SIGMOD International Conference on Management of Data, pp. 93-106, ACM, New York (2008)

[3] Hay, M., Miklau, G., Jensen, D., Towsley, D., Weis, P.: Resisting structural re-identification in anonymized social networks. In: VLDB Endowment, pp. 102-114, ACM, New York (2008)

[4] Zhou, B., Pei, J.: Preserving Privacy in Social Networks Against Neighborhood Attacks. In: IEEE 24th International Conference on Data Engeneering, pp. 506-515, IEEE Press, New York (2008)

[5] Narayanan, A., Shmatikov, V.: De-anonymizing social networks. In: 30th IEEE Symposium on Security and Privacy, pp. 173-187, IEEE Computer Society, Washington (2009)

[6] Narayanan, A., Shi, E., Rubinstein, B.I.P.: Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge. Technical report, Preprint: arXiv:1102.4374v1 (2011)

[7] Singh, L., Zhan, J.: Measuring Topological Anonymity in Social Networks. In: 2007 IEEE International Conference on Granular Computing, pp. 770-770, IEEE Computer Society, Washington (2007)

[8] Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press, Cambridge (1994)

[9] Spertus, E., Sahami, M., Buyukkokten, O.: Evaluating similarity measures: a large-scale study in the orkut social network. In: 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 678-684, ACM, New York (2005)

[10] Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. on Uncertainty, Fuzziness and Knowledge-based System 10 (no. 5), 557–570 (2002)

[11] Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547–579 (1901) (in French)

[12] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z., Barabási, A.-L., Hierarchical organization of modularity in metabolic networks. Science 297, 1551–1555 (2002)

[13] Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley, Reading, MA (1989)

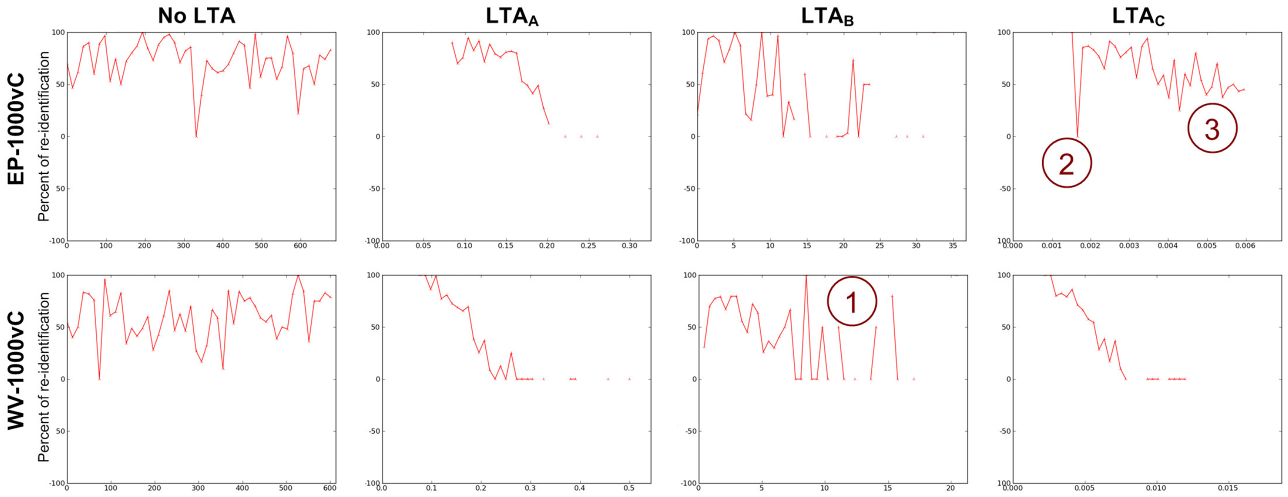

Figure 4. Examples of simulation results for the comparison of LTA variants.

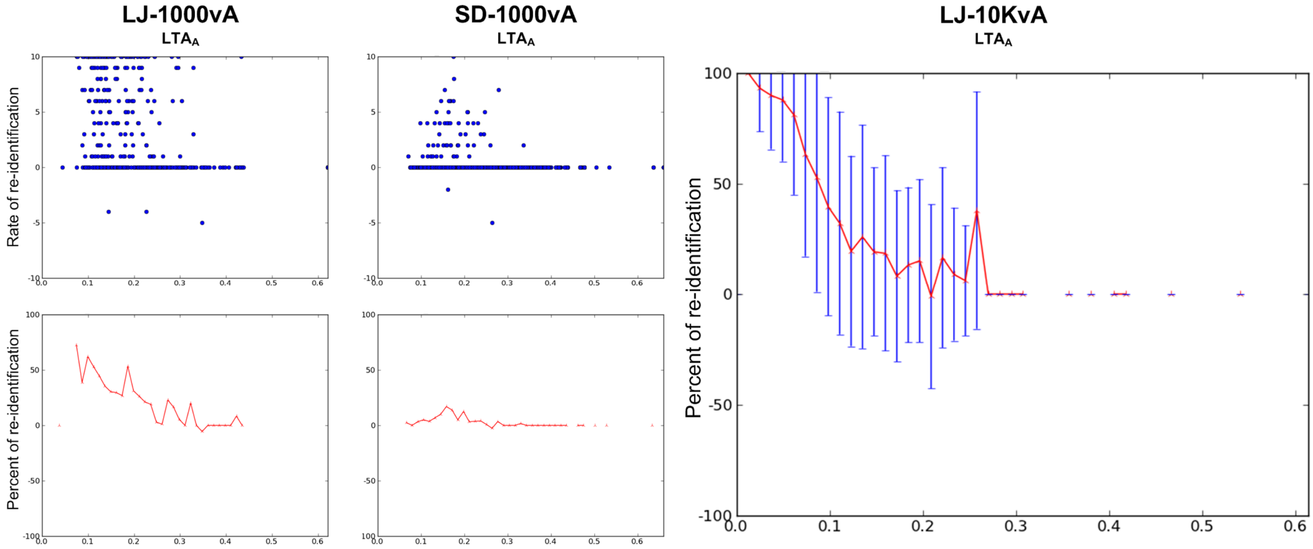

Figure 5. Simulation results on other datasets with LTA_A.