Zero-Knowledge-Private Counting of Group Triangles in Social Networks

Maryam Shoaran[1] and Alex Thomo[2]

[1]Department of Mechatronics, School of Engineering Emerging Technologies, University of Tabriz, Iran
[2]Department of Computer Science, University of Victoria, Canada
Email: mshoaran@tabrizu.ac.ir, thomo@cs.uvic.ca

We introduce a general notion of maturity in social networks that is based on the number of triangles between groups/communities. In order to protect individual privacy upon possible release of such information, we propose privacy mechanisms using zero-knowledge privacy (ZKP), a recently proposed privacy scheme that provides stronger protection than differential privacy (DP) for social-graph data. We present techniques to compute the parameters required to design ZKP methods and finally evaluate the practicality of the proposed methods.

Keywords: Zero-Knowledge Privacy; Differential Privacy; Social Networks; Triangle Count; Randomization Techniques

1. INTRODUCTION

Complex graphs of real world networks have been studied from different aspects. One of the important properties of networks is connectivity, overall or partial. Various measures, such as counting the number of triangles or other simple subgraphs, have been used to characterize connectivity. Connectivity measures can reveal valuable information about networks. For instance, the number of triangles in a graph can indicate how much a graph looks like a social network or how mature a network is (cf. [1]). When a social network is in its early ages pairwise friendship connections are prevalent. As the network grows and activity level increases, individuals increasingly consider establishing friendships with friends of friends, thus forming more triangular connections.

In this paper we consider triangles between groups. More specifically, we are interested in the number of (u, v, w) triangles, where u, v, and w are people in groups g, g', and g'', respectively, and where these groups are not necessarily pair-wise disjoint. This triangle-based measure is quite general. For example, if $g = g' = g''$, the measure gives the number of triangles formed by the members of a single group. Another useful specialization is when $g \cap g' = \emptyset$. Then, our measure gives the bridgeness of g'''s elements with respect to g and g' (see [2]).

Graph measures, similar to other types of aggregate information, are usually released to third parties for different purposes. The release of such information can violate the privacy of individuals in networks. Among the wide range of definitions and schemes presented to protect data privacy, ϵ-Differential Privacy [4, 5, 6] (DP for short) has attracted significant attention in recent years. By adding appropriate noise to the output of a function, DP makes it practically impossible to infer the presence of an individual or a relationship in a database using the released information. While DP stays resilient to many attacks on tabular data, it might not provide sufficient protection in the case of graph data, particularly social networks (cf. [7, 8]). Because of the extensive correlation between the nodes in social networks, not only the participation of a node (or relationship), but also the evidence of such participation has to be protected. This requires a higher level of protection than what DP offers (cf. [8]).

We explain the matter using an example. Suppose there are three groups of nodes g, g', and g'' in a social graph G. We want to publish the number of triangles between these three groups. Let us assume for this example that there is only one person Eve in g''. Suppose that Bob in g is connected to Alice in g', and both are connected to Eve in g'' (a triangle). As a consequence of these relationships, some of Bob's friends make connections to Alice and Eve, thus creating new triangles. What we want to protect is the existence of Bob's connection to Alice (call it the BA-connection). Namely, the privacy of the BA-connection is protected by devising a privacy mechanism that distorts the true answer in a randomized way so that it becomes practically impossible for an attacker to infer whether we have used the original database or the slightly modified version of it where the BA-connection has been removed.

From a counting perspective the existence or not of the BA-connection can change the number of triangles by at most 1. DP works in this case by ensuring that for

any true answer, c or $c-1$, the privatized answer would be pretty much the same. However, this is not strong enough; the existence of Bob's connections influenced the true number of triangles between the three groups not just by 1, but by a bigger number as it caused more triangles to be created by Bob's friends.

In order to provide sufficient data privacy for social graphs, Gehrke, Lui, and Pass proposed "zero-knowledge privacy" (ZKP) in [7]. The definition of ZKP is based on classes of aggregate functions. ZKP guarantees that any additional information that an attacker can obtain about an individual by having access to the sanitized output is indistinguishable from what can be inferred from some sampling-based aggregate.

In ZKP, the sample size k has to be selected wisely. To protect even the evidence of a person's participation the sampling has to be such that, with high probability, very few of the person's neighboring individuals get selected by random sampling.

For instance, suppose in the Bob's example above the network size is 10,000 and the sample size is $\sqrt[3]{10,000^2} = 464$. With such a sampling rate of almost 0.05 the evidence provided by say 10 more triangles caused by Bob's connections will essentially be protected; with a high probability, none of these 10 triangles will be in the sample.

In this paper, we formally define a group-based triangle (GBT) measure in social networks and present a ZKP mechanism to provide connection privacy upon release of such information. Specifically, our first contribution is deriving a general formula to compute the GBT measure, such that it works for any assemblage of node groups. In the second part, we propose methods to compute the sample complexity of the triangle count function. In order to achieve this, we present techniques to express the function as an average of specially designed, synthetic attributes on the nodes of graphs. This is one of the main challenges to be able to use the Hoeffding inequality and the fundamental ZKP proposition (See Proposition 4.1) for GBT measure. Then, we derive precise prescriptions on how to construct ZKP mechanisms for the function.

The rest of the paper is organized as follows. We discuss related work in Section 2. In Section 3, we define our notion of group-based triangles (GBT). Section 4 contains a discussion of the background concepts related to zero-knowledge privacy. In Section 4, we present ZKP mechanisms for releasing GBT measures. We also present our methods to compute the sample complexity of GBT. Section 6 presents a numeric evaluation of the ZKP mechanism, and Section 7 concludes the paper.

2. RELATED WORK

The common goal of privacy preserving methods is to learn from data while protecting sensitive information of the individuals (cf. [9, 10, 11, 12]). k-anonymity for social graphs (cf. [13, 14, 16]) provides privacy by ensuring that combinations of identifying attributes appear at least k times in the dataset. The problem with k-anonymity and other related approaches, e.g. l-diversity [17], is that they assume the adversary has limited auxiliary knowledge. Narayanan and Shmatikov [18] presented a de-anonymization algorithm and claimed that k-anonymity can be defeated by their method using auxiliary information accessible by the adversary.

Among a multitude of different techniques, differential privacy (DP) [4, 5, 19, 20] has become one of the leading methods to provide individual privacy. Various differentially private algorithms have since been developed for different domains, including social networks [21, 22]. However, DP can suffer in social networks where specific auxiliary information, such as graph structure and friendship data, is easily available to the adversary. Important works showing the shortcomings of DP are [8, 23, 3].

Gehrke, Lui, and Pass in [7] present the notion of zero-knowledge privacy that is appealing for achieving privacy in social networks. Zero-knowledge privacy (ZKP) guarantees that what can be learned from a dataset including an individual is not more than what can be learned from sampling-based aggregates computed on the dataset without that individual.

Works [2, 24] use ZKP to release connectedness and bridgeness statistics in social networks. They are different from the current work, where we aim at privately releasing group-based triangle counts for social networks. Specifically, [24] is about ZKP-Graph Summarization (ZKP-GS). It considers some statistics in graphs which involve edges between one or two groups. If intuitively we assume $\#edges(g, g') > \#triangles(g, g', g'')$, in a similar circumstance this results in a smaller noise scale. Furthermore, ZKP-GS has different techniques than current paper to compute the sample complexity of corresponding functions.

3. GRAPHS, GROUPS, AND TRIANGLES

We denote a graph as $G = (V, E)$, where V is the set of nodes and $E \subseteq V \times V$ is the set of edges connecting the nodes. We consider $\mathcal{S} \subset 2^V$ to be a set of node groups of size r or more that a social network wants to release statistics about. Let g, g', and g'' be three groups in $\mathcal{S}$.

Definition 3.1. *The* group-based triangle (GBT) *measure of g, g', and g'' is defined as*

$$GBT(g, g', g'') = \frac{|\{(u,v,w):\ u\in g, v\in g', w\in g'',\ \{(u,v),(u,w),(v,w)\}\subseteq E\}|}{\max|\{(u,v,w):\ u\in g, v\in g', w\in g''\}|}$$

Intuitively, $GBT(g, g', g'')$ is the fraction of the number of actual triangles between the nodes of different groups over the number of all possible such triangles. Throughout the paper, we will refer to

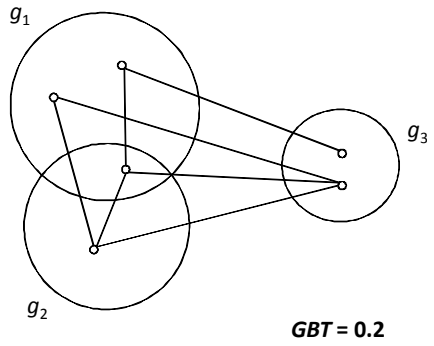

FIGURE 1. Group-based triangle count

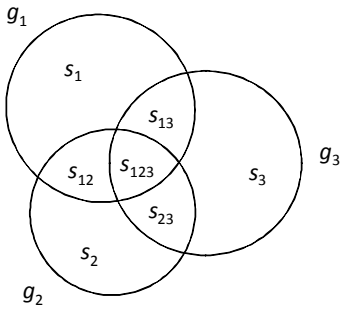

FIGURE 2. Disjoint subsets

$GBT(g, g', g'')$ as GBT whenever g, g', and g'' are clear from the context.

The definition of GBT is quite general. It contains as special cases simpler triangle-based measures used in literature. For example, by setting $g = g' = g'' = V$, we obtain the total number of triangles in the network (cf. [1, 25, 26, 27]).

To facilitate the exposition and derivation of some equations, we will also use occasionally g_1, g_2, g_3 to refer to groups. This is the case for the rest of this section.

Example 1. Fig. 1 shows a graph G with three groups g_1, g_2, and g_3, having three, two, and two nodes, respectively. There are three edges connecting the nodes of g_1 and g_2, and four edges connecting g_3 to g_1 and g_2. These edges form two triangles in total between groups g_1, g_2, and g_3. The number of all possible triangles between three groups is 10 (See Proposition 3.1). Therefore, we have $GBT = 0.2$.

We derive in the following a formula to compute the maximum number of triangles that can exist between any three groups. The set of all valid triangles is denoted by L, where each triangle is represented by a node triplet (u, v, w). $|L|$ is the denominator in the definition of GBT.

Let g_1, g_2, and g_3 be three groups with possible intersections. We consider the following notation, as illustrated in Fig. 2, for all possible disjoint subsets created by the intersections between the groups.

$$\begin{aligned} s_i &= \{v | v \in g_i, v \notin g_j, \text{ and } v \notin g_k\} \\ s_{ij} &= \{v | v \in g_i, v \in g_j, \text{ and } v \notin g_k\} \\ s_{ijk} &= \{v | v \in g_i, v \in g_j, \text{ and } v \in g_k\} \end{aligned}$$

where $i, j, k \in \{1, 2, 3\}$ and $i \neq j$, $j \neq k$, and $i \neq k$. Note that the subset indices are commutative, e.g., $s_{ij} = s_{ji}$. With slight abuse of notation, we will use s_{ij} to denote the size of the s_{ij} set as well.

Proposition 3.1. *The maximum number of triangles between three node groups is computed by the equation in Table 1.*

Term (1a) counts the number of triangles having at least two vertices in *only* one group (each to a separate group). That is, they are not in the areas of intersection with other groups. The third vertex has to be in any of the subsets of the third group, including the intersections with the other two groups. The coefficients and the subtraction take care of the repetitions occurring in the summation.

Term (1b) considers the triangles having at least two vertices in distinctive intersections of any two groups. Since these two vertices can be assumed to be in any of the three groups, the third vertex can be in any of the other subsets.

Term (1c) gives the number of possible triangles in which one vertex is in only one group, the other vertex is in the intersection of any two groups, and the third one is in all three groups.

Two vertices of a triangle can both be placed in the intersection of any two groups. Such triangles are counted by term (1d). The third vertex has to be in any of the subsets of the third group. Similarly, (1e) counts the number of triangles with two vertices in s_{123}, the intersection of all three groups.

As vertices in s_{123} belong to all groups, triangles can reside entirely, with all three vertices, inside this subset, counted by (1f).

Example 2. In Fig. 1 the total number of possible triangles is computed as follows. We have $s_1 = 2$, $s_2 = 1$, $s_3 = 2$, $s_{12} = 1$, and $s_{13} = s_{23} = s_{123} = 0$. Therefore, the only term that is not zero is (3.1) computed as the following.

$$\begin{aligned} |L| &= (s_1 \cdot s_2 \cdot (s_3 + s_{13} + s_{23} + s_{123})) + \\ &\quad (s_1 \cdot s_3 \cdot (s_2 + s_{12} + s_{23} + s_{123})) + \\ &\quad (s_2 \cdot s_3 \cdot (s_1 + s_{12} + s_{13} + s_{123})) - 2 \cdot s_1 \cdot s_2 \cdot s_3 \\ &= (2 \cdot 1 \cdot 2) + (2 \cdot 2 \cdot (1 + 1)) + \\ &\quad (1 \cdot 2 \cdot (2 + 1)) - 2 \cdot 2 \cdot 1 \cdot 2 = 10 \end{aligned}$$

4. BACKGROUND ON ϵ-ZERO-KNOWLEDGE PRIVACY

Zero-Knowledge Privacy (ZKP), introduced by [7], is an enhanced privacy scheme that guarantees stronger

$$
\begin{aligned}
|L| &= \left(\sum_{\substack{i,j,k\in\{1,2,3\}\\ i<j,i\neq j,j\neq k,i\neq k}} (s_i \cdot s_j)(s_k + s_{ik} + s_{jk} + s_{ijk})\right) - 2s_1 \cdot s_2 \cdot s_3 && \text{(1a)}\\
&+ \frac{1}{2}\left(\sum_{\substack{i,j,k\in\{1,2,3\}\\ i\neq j,j\neq k,i\neq k}} (s_{ij} \cdot s_{jk})(s_i + s_j + s_k + s_{ik} + s_{ijk})\right) - 2s_{12} \cdot s_{13} \cdot s_{23} && \text{(1b)}\\
&+ \left(\sum_{i\in\{1,2,3\}} s_i\right) \cdot \left(\sum_{\substack{j,k\in\{1,2,3\}\\ j<k}} s_{jk}\right) \cdot s_{123} && \text{(1c)}\\
&+ \sum_{\substack{i,j,k\in\{1,2,3\}\\ i<j,i\neq j,j\neq k,i\neq k}} \binom{s_{ij}}{2}(s_k + s_{ik} + s_{jk} + s_{ijk}) && \text{(1d)}\\
&+ \binom{s_{123}}{2}(s_1 + s_2 + s_3 + s_{12} + s_{13} + s_{23}) && \text{(1e)}\\
&+ \binom{s_{123}}{3} && \text{(1f)}
\end{aligned}
$$

TABLE 1. Maximum number of possible triangles between three groups

privacy protection, compared to other currently well-known methods such as differential privacy (DP), especially in social networks. Due to the extensive influence in such networks, the presence of a single element (node or connection) can lead to the creation of several new elements in the network. Therefore, in such a setting a privacy mechanism needs to protect not only the participation of an element in the network, but also the evidence of such a participation, i.e. the presence of new elements created under the influence of the element in focus.

ZKP requires that whatever an intelligent agent (*adversary*) can discover from the sanitized output of the mechanism is not more than what can be discovered by an equally gifted agent that only has access to some sample-based aggregate information. The latter agent is sometimes referred as *simulator*.

Let G be a graph. We denote by G_{-*} a graph obtained from G by removing a piece of information (for example an edge). G and G_{-*} are called *neighboring graphs*.

Let M be the privacy mechanism that securely releases the answer to a query on graph G, and let A be the intelligent agent that operates on output $M(G)$, that is, privatized answer, trying to breach the privacy of some individual. Let S be a simulator as capable as A, that would have access to some aggregate information obtained by an algorithm $T \in agg$. Note that, the assumed algorithm T only would compute *approximate* answers to aggregate functions by sampling graph G_{-*}, i.e. the graph that misses the piece of information which should be protected.

Definition 4.1. *(Zero-Knowledge Privacy* [7]*) The mechanism M is* **ϵ-zero-knowledge private with respect to** *agg if there exists a $T \in agg$ such that for every adversary A, there exists a simulator S such that for every G, every $z \in \{0,1\}^*$, and every $W \subseteq \{0,1\}^*$, the following hold:*

$$
\begin{aligned}
Pr[A(M(G), z) \in W] &\leq e^{\epsilon} \cdot Pr[S(T(G_{-*}), z) \in W]\\
Pr[S(T(G_{-*}), z) \in W] &\leq e^{\epsilon} \cdot Pr[A(M(G), z) \in W]
\end{aligned}
$$

where probabilities are taken over the randomness of M and A, and T and S.

This definition assumes that both the adversary and simulator have access to some general and easily accessible auxiliary information z, such as graph structures or the groups the individuals belong in.

Note that, based on the application settings the selection of k –the sample size– in agg algorithms is very important. It should be chosen so that with high probability very few of the elements (nodes or edges) related with the element whose information has to be private will be chosen. We will often index agg by k as agg_k to stress the importance of k. To satisfy the ZKP definition, a mechanism should use $k = o(n)$, say $k = \sqrt{n}$ or $k = \sqrt[3]{n^2}$, where n, the number of nodes in the database, is sufficiently large (see [7]). DP is a special case of ZKP where $k = n$.

Achieving ZKP. Let $f : \mathbf{G} \to \mathbb{R}^m$ be a function that produces a vector of length m from a graph database. For example, given graph G, and the set of node groups $\mathcal{S}$, f produces GBT for m triplets of groups. We consider the L_1-Sensitivity to be defined as follows.

Definition 4.2. *(L_1-Sensitivity) For $f : \mathbf{G} \to \mathbb{R}^m$, the L_1-sensitivity of f is*

$$\Delta(f) = \max_{G', G''} ||f(G') - f(G'')||_1$$

for all neighboring graphs G' and G''.

Another essential definition is that of "sample complexity".

Definition 4.3. *(Sample Complexity* [7]*) A function $f : Dom \to \mathbb{R}^m$ is said to have (δ, β)-**sample***

complexity *with respect to agg if there exists an algorithm $T \in agg$ such that for every $D \in Dom$ we have*

$$Pr[||T(D) - f(D)||_1 \leq \delta] \geq 1 - \beta.$$

T is said to be a (δ, β)-sampler for f with respect to agg.

This definition bounds the probability of error between the randomized computation (approximation) of function f and the expected output of f. Functions with low sample complexity (smaller δ and β) can be computed more accurately using random samples from the input data.

When the released information, as typical, is real numbers, the ZKP mechanism *San* achieves the privacy by adding noise to each of the numbers independently.

Let $Lap(\lambda)$ be the zero-mean Laplace distribution with scale λ, and variance $2\lambda^2$. The scale of Laplace noise in ZKP is properly calibrated to the sample complexity of the function that is to be privately computed. The following proposition expresses the relationship between the sample complexity of a function and the level of zero knowledge privacy achieved by adding Laplace noise to the outputs of the function.

Proposition 4.1. ([7]) *Suppose $f : \mathbf{G} \rightarrow [a, b]^m$ has (δ, β)-sample complexity with respect to agg. Then, mechanism*

$$San(G) = f(G) + (X_1, \ldots, X_m),$$

where $G \in \mathbf{G}$, and $X_j \backsim Lap(\lambda)$ for $j = 1, \cdots, m$ independently, is

$$\ln\left((1-\beta)e^{\frac{\Delta(f)+\delta}{\lambda}} + \beta e^{\frac{(b-a)m}{\lambda}}\right)$$

–ZKP with respect to agg.

5. ZKP MECHANISM FOR GBT MEASURE

In this section we design an ϵ-ZKP mechanism to privately release GBT measures. Let f be the function that given graph G and set $\mathcal{S}$ produces a c-dimensional vector of GBT measures (numbers), where $c \leq \binom{|\mathcal{S}|}{3}$.

Let $f = [f_1, \ldots, f_t]$ be the vector that is to be privately released. We apply a separate San_i (ZKP) mechanism, for $i \in [1, t]$, to each of the elements of f. Let us assume that each San_i provides ϵ_i-ZKP for f_i with respect to agg_{k_i}, where $k_i = k(n)/t$ and $n = |V|$. Then, based on the following proposition, f will be $\left(\sum_{i=1}^{t} \epsilon_i\right)$-ZKP with respect to $agg_{k(n)}$, where $k(n) = \sum_{i=1}^{t} k_i$.

Proposition 5.1. *(Sequential Composition [7]) Suppose San_i, for $i \in [1, n]$, is an ϵ_i-ZKP mechanism with respect to agg_{k_i}. Then, the mechanism resulting from composing[3] San_i's is $(\sum_{i=1}^{n} \epsilon_i)$-ZKP with respect to $agg_{(\sum k_i)}$.*

Consider G and G_{-e}, where G_{-e} is a neighboring graph of G obtained from G by removing edge e. The goal of our mechanism is to protect the privacy of the connections between the nodes of different groups upon release of GBT counts. Therefore, we assume that the removed edge e is an edge between two nodes of two different groups in $\mathcal{S}$. To compute the sensitivity of GBT measures, we consider two extreme cases; first the case when the three node groups are pairwise disjoint, and second when they are identical. In the first case, removing an edge between two node groups g' and g'' can change by at most $|g|$ the numerator of $GBT(g, g', g'')$ in G_{-e}. Hence, the sensitivity of the GBT function in this case is $\Delta_1(GBT) = \max \frac{|g|}{|g| \cdot |g'| \cdot |g''|} = \frac{1}{r^2}$ where r is the minimum group size in $\mathcal{S}$. Note that the denominator is the maximum number of valid triangles when all three groups are disjoint.

In the second case when the groups are identical, $g = g' = g''$, the sensitivity is $\Delta_2(GBT) = \max \frac{|g|-2}{\binom{|g|}{3}} = \frac{6}{r(r-1)}$. Therefore, the overall sensitivity of GBT function is $\Delta(GBT) = \frac{6}{r(r-1)}$.

Now, suppose $GBT(g, g', g'')$ is an element of f, where g, g', and g'' are groups in $\mathcal{S}$. Let $San = GBT(g, g', g'') + Lap(\lambda)$ be a ZKP mechanism which adds random noise selected from $Lap(\lambda)$ distribution to the output of $GBT(g, g', g'')$ in order to achieve ZKP. Our goal here is to come up with the right λ to achieve a predefined level of ZKP.

Based on the definition of ZKP, one should first know the sample complexity of the GBT function. For this, without any change in semantics, we will express GBT so that it computes an average rather than a fraction of two counts. Then, using the *Hoeffding* inequality (cf. [28]) we compute the sample complexity of GBT.

Expressing GBT**.** In addition to regular node attributes (if any), we introduce $\binom{|\mathcal{S}|}{2}$ new boolean attributes, one for each group pair in $\mathcal{S}$. We denote each new attribute by upper-case I indexed by a group-pair id. Each attribute $I_{g_i g_j}$ (I_{ij} for short) is a boolean vector of dimension $|g_i| \cdot |g_j|$, where each dimension corresponds to a possible edge between g_i and g_j. A node u in graph $G(V, E)$ will have $I_{ij}(u)[vw] = 1$, where $v \in g_i$ and $w \in g_j$, if $\{(u, v), (u, w), (v, w)\} \subseteq E$, and $I_{ij}(u)[vw] = 0$, otherwise. Note that, we have $u \neq v$, $v \neq w$, and $w \neq u$. For each triplet of groups g_i, g_j, and g_k we can verify that:

[3] A set of computations that are separately applied on *one* database and each provides ZKP in isolation, also provides ZKP for the set. In the case when the output of the computations is not independent from each other, the composition is called sequential (as opposed to parallel composition).

Proposition 5.2.

$$\begin{aligned} GBT(g_i,g_j,g_k) &= \frac{\sum_{\substack{v\in g_i,w\in g_j,u\in g_k\\(u,v,w)\in L}} I_{ij}(u)[vw]}{|L|} \\ &= \frac{\sum_{\substack{v\in g_j,w\in g_k,u\in g_i\\(u,v,w)\in L}} I_{jk}(u)[vw]}{|L|} \\ &= \frac{\sum_{\substack{v\in g_k,w\in g_i,u\in g_j\\(u,v,w)\in L}} I_{ki}(u)[vw]}{|L|} \end{aligned}$$

Therefore, the $GBT(g_i,g_j,g_k)$ measure can be viewed as the average of $I_{ij}[.]$, or $I_{jk}[.]$, or $I_{ki}[.]$ over the nodes of g_k, g_i, g_j, respectively.

ZKP Mechanism. Let $G=(V,E)$ be a graph enriched with boolean attributes as explained above. We would like to determine the value of $\lambda>0$ for the $Lap(\lambda)$ distribution which will be used to add random noise to $GBT(g,g',g'')$ included in f. For this, first we compute the sample complexity of GBT to be able to use Proposition 4.1 and establish an appropriate value for λ.

Let T be a randomized algorithm in agg_k, the class of randomized algorithms that operates on an input graph G. To randomly sample a graph G, algorithm T would uniformly select $k=k(n)/t$ random nodes from V, read their attributes, and retrieve all edges[4] incident to these k sample nodes.[5]

With this sampling, the nodes in the groups of $\mathcal{S}$ and the edges between them would be randomly sampled as well. Let us assume that we have a sample of each group and edges between groups, and a sample of group g is denoted as g_k. Then, algorithm T would approximate GBT using sampled graph data. For the sample complexity of $GBT(g,g',g'')$, since we expressed it as averages, we can use the Hoeffding inequality as follows;

$$Pr[|T(g,g',g'')-GBT(g,g',g'')|\le\delta] \quad\ge\quad 1-2e^{-2|L_k|\delta^2}$$

where $|L_k|$ is the number of all possible valid triangles between sample groups g_k, g'_k, and g''_k.

From this and Definition 4.3, we have that the GBT function has $\left(\delta, 2e^{-2|L_k|\delta^2}\right)$-sample complexity with respect to agg_k.

Now we make the following substitutions in the formula of Proposition 4.1: $\beta = 2e^{-2|L_k|\delta^2}$, $\Delta(GBT(g,g',g'')) = \frac{6}{r(r-1)}$, $b-a=1$, and $m=1$. From this, mechanism San becomes

$$\ln\left(e^{\frac{\frac{6}{r(r-1)}+\delta}{\lambda}} + 2e^{\frac{1}{\lambda}-2|L_k|\delta^2}\right) \text{ -ZKP}$$

with respect to agg_k.

[4]Clearly, only non-dangling incident edges, whose both end nodes have been sampled, will be retrieved.

[5]For other possible methods of graph sampling see for example [7].

Similarly to DP, we set λ, the Laplace noise scale, to be proportional to "the error" as can be measured in the ZKP method by the sum of the sensitivity $\Delta(GBT)$ and sampling error δ, and inversely proportional to the ZKP privacy level. Regarding δ, we can consider for instance $\delta=\frac{1}{\sqrt[3]{|L_k|}}$, and hence,

$$\lambda=\frac{\Delta(GBT)+\delta}{\epsilon}=\frac{1}{\epsilon}\left(\frac{6}{r(r-1)}+\frac{1}{\sqrt[3]{|L_k|}}\right)$$

From all the above, the privacy level obtained will be[6]

$$\begin{aligned} &\ln\left(e^{\frac{\frac{6}{r(r-1)}+\delta}{\lambda}} + 2e^{\frac{1}{\lambda}-2|L_k|\delta^2}\right) \\ =\ &\ln\left(e^{\epsilon} + 2e^{\frac{\epsilon}{6/r(r-1)+1/\sqrt[3]{|L_k|}}-2\sqrt[3]{|L_k|}}\right) \\ \le\ &\ln\left(e^{\epsilon}+2e^{-\sqrt[3]{|L_k|}}\right) \\ \le\ &\epsilon+2e^{-\sqrt[3]{|L_k|}}. \end{aligned}$$

Thus, we have that by adding noise randomly selected from $Lap\left(\frac{1}{\epsilon}\left(\frac{6}{r(r-1)}+\frac{1}{\sqrt[3]{|L_k|}}\right)\right)$ distribution to GBT, San will be $\left(\epsilon+2e^{-\sqrt[3]{|L_k|}}\right)$-ZKP with respect to agg_k.

Example 3. Let graph G be a social graph with ten million participants/nodes ($|V|=n=10,000,000$), and g, g', and g'' be three node groups in $\mathcal{S}$. Suppose that the minimum group size in $\mathcal{S}$ is $r=100$, and we would like to report $GBT(g,g',g'')$. To privately release GBT measures, a randomized algorithm T would uniformly select k nodes and approximate the value of $GBT(g,g',g'')$ using sample data.

The actual value of function $GBT(g,g',g'')$ is computed on G. Suppose that the number of all possible valid triangles between group samples g_k, g'_k, and g''_k is $|L_k|=300,000$. Let (δ,β) be the sample complexity of $GBT(g,g',g'')$ where

$$\delta=\frac{1}{\sqrt[3]{|L_k|}}=\frac{1}{\sqrt[3]{300,000}}=0.0149.$$

$$\beta=2e^{-2|L_k|\delta^2}=2e^{-2*(300,000)*(0.0149)^2}=2.82*10^{-58}.$$

The sensitivity of GBT is

$$\Delta(GBT)=\frac{6}{r(r-1)}=\frac{6}{100*99}=0.0006.$$

Now, if we would like to use a mechanism which is 0.1-ZKP, we can add random noise selected from a Laplace distribution with scale

$$\lambda=\frac{\Delta(GBT)+\delta}{\epsilon}=\frac{0.0006+0.0149}{0.1}=0.155$$

[6]Note that the inequality is true because ϵ is a small number.

to the actual value of $GBT(g, g', g'')$. With this noise scale, the ZKP privacy level of the mechanism is precisely

$$\epsilon \leq \left(\epsilon + 2e^{-\sqrt[3]{|L_k|}}\right) = \left(0.1 + 2 * e^{-66.94}\right) \approx 0.1$$

with respect to agg_k.

To compare this result with Differential Privacy (DP) method in [4, 20], in order to get 0.1-DP privacy, we would have $\lambda = \Delta(GBT)/0.1 = 0.0006$. As it is clear from the computation, the noise scale in DP method is quite smaller. While the ZKP noise is sufficient to protect all the evidence of participation.

6. EVALUATION

In our methods, the amount of noise added to the output is independent of the database, and it only depends on the function we compute and their sensitivities. Therefore, the following analysis is valid for any database.

6.1. Parameters Affecting Noise Scale

Sampling error δ is an important factor specifying λ based on the formula of noise scale $\lambda = \frac{\Delta(f)+\delta}{\epsilon}$. The error in turn has reverse connection with the size of group samples and therefore, with the sample size and size of the database graph. Recall that throughout the paper we considered the error to be $\delta = \frac{1}{\sqrt[3]{|L_k|}}$.

Fig. 3 illustrates the relationship between the noise scale λ and the parameter $|L_k|$. In this figure we assumed that the minimum group size is $r = 100$, and the ZKP-level ϵ is 0.1. The figure shows that as parameter $|L_k|$ decreases from five hundred thousand to one thousand the noise scale increases non-linearly to the amounts that are not practical in our setting. Therefore, our proposed ZKP mechanism is perfect for big databases with large sample sizes. However, even $|L_k| = 500,000$ implies some sample group sizes around $k = 80$, which is reasonable in social graphs with only millions of participants, provided that the groups include some linchpin nodes that build a dense subgraph. (recall that $|L_k|$ is the number of triangles between group samples). Hence, we conclude that the proposed ZKP mechanism works well with small dense data graphs as well as large graphs.

6.2. The Noise

The analysis in this section aims to provide a better understanding of the amount of noise added to outputs. The cumulative distribution function of Laplace distribution in an interval $[-z, z]$ is computed as follows,

$$Pr(-z \leq x \leq z) = \int_{-z}^{z} \frac{1}{2\lambda} e^{\frac{-|x|}{\lambda}} dx = 1 - e^{\frac{-z}{\lambda}}.$$

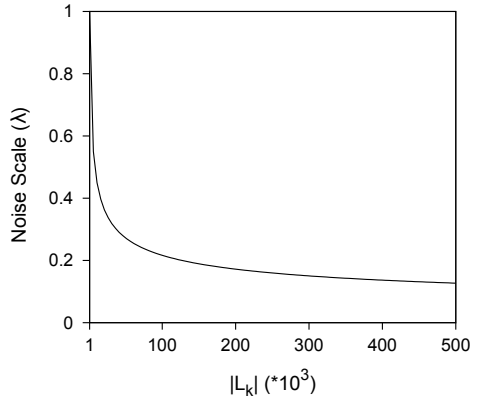

FIGURE 3. Relationship between noise scale and sample size.

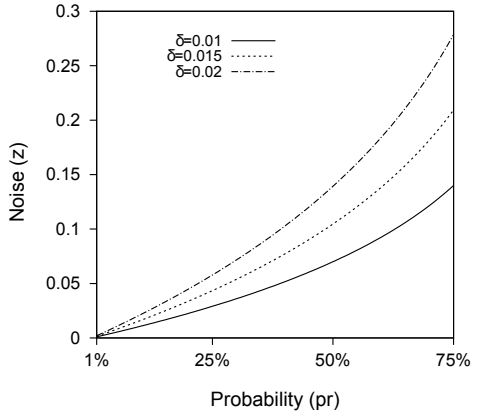

FIGURE 4. Probability vs noise.

Let $pr = Pr(|x| \leq z)$. Value z for a specified cumulative probability pr can be calculated using the above equation as

$$z = -\lambda \cdot \ln(1 - pr) = -\frac{\Delta(f) + \delta}{\epsilon} \cdot \ln(1 - pr).$$

Figure 4 illustrates the maximum absolute noise z as a function of cumulative probability pr for three different values of δ when $\epsilon = 0.1$ and $\Delta(f) = 0.0001$. Each point (pr, z) on the curve for a given δ means that

> pr percent of the time the random noise has an absolute value of at most z.

For example, for $\delta = 0.02$ we have that 50% of the time the absolute value of noise is at most 0.14, and 75% of the time it is at most 0.28.

7. CONCLUSIONS

We addressed zero-knowledge private methods for releasing group-based triangle (GBT) measures for social networks. The application of our technique is crucial in order to have a secure public release of such graph measures. We introduced methods to compute the ZKP parameters, specifically the sample complexity. We showed that the proposed technique

is practically useful for large as well as small dense data graphs. This is different from the mechanism presented in [24], which is useful only for very large social graphs. As future work we aim to generalize the notion of group-based triangles to other inter-group patterns expressed by queries on edge-labeled graphs ([29, 30, 31]). Also, we plan to extend our results to graphs with probabilities in their edges [32].

REFERENCES

[1] Rajaraman, A., Leskovec, J., and Ullman, J. D. (2010) *Mining of Massive Datasets*. Stanford University.

[2] Shoaran, M. and Thomo, A. (2014) Zero-knowledge private computation of node bridgeness in social networks. *Advanced Information Systems Engineering Workshops - CAiSE 2014 International Workshops, Thessaloniki, Greece, June 16-20, 2014. Proceedings*, pp. 31–43.

[3] Shoaran, M. and Thomo, A., Weber, H. J.: (2012) Differential Privacy in Practice. *Secure Data Management*, pp. 14–24.

[4] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. *TCC*, pp. 265–284.

[5] Dwork, C. (2008) Differential privacy: A survey of results. *TAMC*, pp. 1–19.

[6] Dwork, C. (2010) Differential privacy in new settings. *SODA*, pp. 174–183.

[7] Gehrke, J., Lui, E., and Pass, R. (2011) Towards privacy for social networks: A zero-knowledge based definition of privacy. *TCC*, pp. 432–449.

[8] Kifer, D. and Machanavajjhala, A. (2011) No free lunch in data privacy. *SIGMOD Conference*, pp. 193–204.

[9] Shen, Y. and Thonnard, O. (2014) MR-TRIAGE: scalable multi-criteria clustering for big data security intelligence applications. *2014 IEEE International Conference on Big Data, Big Data 2014, Washington, DC, USA, October 27-30, 2014*, pp. 627–635.

[10] Mohammed, N., Chen, R., Fung, B. C. M., and Yu, P. S. (2011) Differentially private data release for data mining. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pp. 493–501.

[11] Xu, J., Zhang, Z., Xiao, X., Yang, Y., and Yu, G. (2012) Differentially private histogram publication. *IEEE 28th International Conference on Data Engineering (ICDE 2012), Washington, DC, USA (Arlington, Virginia), 1-5 April, 2012*, pp. 32–43.

[12] Friedman, A. and Schuster, A. (2010) Data mining with differential privacy. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pp. 493–502.

[13] Liu, K. and Terzi, E. (2008) Towards identity anonymization on graphs. *SIGMOD Conference*, pp. 93–106.

[14] Chester, S., Kapron, B. M., Ramesh, G., Srivastava, G., Thomo, A., and Venkatesh, S. (2011) k-anonymization of social networks by vertex addition. *ADBIS*, pp. 107–116.

[15] Chester, S., Kapron, B. M., Ramesh, G., Srivastava, G., Thomo, A., and Venkatesh, S. (2013) Why waldo befriended the dummy? k-anonymization of social networks with pseudo-nodes. *Social Netw. Analys. Mining*, **3**, 381–399.

[16] Chester, S., Kapron, B. M., Ramesh, G., Srivastava, G., Thomo, A., and Venkatesh, S. (2013) Why waldo befriended the dummy? k-anonymization of social networks with pseudo-nodes. *Social Netw. Analys. Mining*, **3**, 381–399.

[17] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007) *L*-diversity: Privacy beyond *k*-anonymity. *TKDD*, **1**.

[18] Narayanan, A. and Shmatikov, V. (2009) De-anonymizing social networks. *IEEE Symposium on Security and Privacy*, pp. 173–187.

[19] Blum, A., Dwork, C., McSherry, F., and Nissim, K. (2005) Practical privacy: the sulq framework. *PODS*, pp. 128–138.

[20] Dwork, C. (2006) Differential privacy. *ICALP (2)*, pp. 1–12.

[21] Hay, M., Li, C., Miklau, G., and Jensen, D. (2009) Accurate estimation of the degree distribution of private networks. *ICDM*, pp. 169–178.

[22] Rastogi, V., Hay, M., Miklau, G., and Suciu, D. (2009) Relationship privacy: output perturbation for queries with joins. *PODS*, pp. 107–116.

[23] Kifer, D. and Machanavajjhala, A. (2012) A rigorous and customizable framework for privacy. *PODS*, pp. 77–88.

[24] Shoaran, M., Thomo, A., and Weber-Jahnke, J. H. (2013) Zero-knowledge private graph summarization. *BigData Conference*, pp. 597–605.

[25] Schank, T. and Wagner, D. (2005) Finding, counting and listing all triangles in large graphs, an experimental study. *Experimental and Efficient Algorithms, 4th InternationalWorkshop, WEA 2005, Santorini Island, Greece, May 10-13, 2005, Proceedings*, pp. 606–609.

[26] Tsourakakis, C. E., Kang, U., Miller, G. L., and Faloutsos, C. (2009) DOULION: counting triangles in massive graphs with a coin. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pp. 837–846.

[27] Suri, S. and Vassilvitskii, S. (2011) Counting triangles and the curse of the last reducer. *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pp. 607–614.

[28] Mitzenmacher, M. and Upfal, E. (2005) *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA.

[29] Grahne, G., Thomo, A. (2003) Algebraic rewritings for optimizing regular path queries. *Theoretical Computer Science*, 296 (3), pp. 453–471.

[30] Stefanescu, C. D., Thomo, A. Thomo, L. (2005) Distributed evaluation of generalized path queries. *Proceedings of the 2005 ACM symposium on Applied Computing (SAC 2005)*, pp. 610–616.

[31] Grahne, G., Thomo, A., Wadge, W. (2007) Preferentially annotated regular path queries. *Proceedings of*

the 11th International Conference on Database Theory (ICDT 2007), Barcelona, Spain, January 10-12, 2007, pp. 314–328.

[32] Nasrin Hassanlou, Maryam Shoaran, Alex Thomo. (2013) Probabilistic Graph Summarization. *Proceedings of WAIM 2013*, pp. 545–556.