Zero-Knowledge Private Computation of Node Bridgeness in Social Networks

Maryam Shoaran[1] and Alex Thomo[2]

[1] University of Tabriz, Tabriz, Iran
mshoaran@tabrizu.ac.ir
[2] University of Victoria, Victoria, Canada
thomo@cs.uvic.ca

Abstract. We introduce a bridgeness measure to assess the influence of a node in the connectivity of two groups (communities) in a social network. 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.

1 Introduction

For many years, complex graphs of real world networks have been studied from different aspects. One major line of research is devoted to the study of the role of nodes and edges in the functionality and structure of networks. Various indices have been proposed to characterize the significance of nodes and edges. Centrality measures like *degree*, *closeness*, and *betweenness* (cf. [30, 12] are used to determine the role of a node in maintaining the overall and partial connectivity of networks. Various definitions of *bridgeness* are proposed to measure the role of nodes or edges [28, 5]. Here we define another notion of bridgeness to measure the effect of a node (particularly a linchpin[3]) on the connectivity of two groups (communities) in a social graph.

Graph characteristics like bridgeness, similar to other aggregate information, are usually released to the 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 [11, 9, 10] (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

[3] Highly active members of networks usually act as linchpins. For example, highly active authors or actors in collaborative networks play an essential role in connecting sub-units (*communities* or *clusters*) [25].

attacks on tabular data, it might not provide sufficient protection in the case of graph data, particularly social networks (c.f.[13, 19]). Because of the extensive correlation between the nodes in social networks, not only the participation of a node (or relationship), but also the evidences of such participation have to be protected. And this requires a higher level of protection than DP (cf. [19]).

We explain the matter using an example. Suppose there are two groups of nodes g_1, g_2, and a node p in a social graph G. We want to publish the number of triangles between these three disjoint components of G. Suppose that there is a triangle between Bob in g_1, Alice in g_2, and p. As a consequence of such relationship, some friends of Bob make connections to Alice and to p, thus creating new triangles. What we want to protect is Bob's edge to Alice. From a counting perspective the existence or not of this edge can change the answer by 1. DP works in this case by ensuring that for any true answer, c or $c-1$, the sanitized answer would be pretty much the same. However, this is not strong enough; the existence of Bob's edge influenced the true number of triangles not just by 1, but by a bigger number as it caused more triangles to be created by Bob's and Alice's friends.

In order to provide sufficient data privacy for social graphs, Gehrke, Lui, and Pass proposed "zero-knowledge privacy" (ZKP) in [13]. 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 privatized output is indistinguishable from what can be inferred from some sampling-based (approximate) aggregates. The level of privacy in ZKP mechanisms is defined using the sample complexity of aggregates. For instance, suppose in the Bob's example above the network size is 10000 and the sample size is $\sqrt[3]{10000^2} = 464$. 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 use ZKP to provide connection privacy when releasing inter-community bridgeness of linchpin nodes. We define a natural notion of bridgeness in social graphs and present a ZKP mechanism for private release of bridgeness. Specifically, we propose methods to compute the sample complexity of the bridgeness function. In order to achieve this, we present techniques to express the function as averages of specially designed, synthetic attributes on the nodes of graphs. 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 bridgeness. Section 4 contains an elaborate discussion of the background concepts related to zero-knowledge privacy. In Section 4, we present ZKP mechanisms for bridgeness measure. Also in this section, we present our methods to compute the sample complexity of bridgeness. Section 6 presents a numeric evaluation of the ZKP mechanism, and Section 7 concludes the paper.

2 Related Work

Massive networks, graphs, and graph databases have become very popular for more than a decade (c.f. [14, 2, 35, 3, 15, 31, 32, 4]). Computing statistics and summarizations for graph data is very important as it is difficult to understand their structure using other means (c.f. [36, 39, 18, 37, 38, 16, 22]).

The common goal of privacy preserving methods is to learn from data while protecting sensitive information of the individuals. k-anonymity for social graphs (cf. [23, 6, 21, 7]) 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 [24], is that they assume the adversary has limited auxiliary knowledge. Narayanan and Shmatikov [27] 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) [1, 8, 11, 9] 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 [17, 29]. However as already shown, 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 [19, 20].

Gehrke, Lui, and Pass in [13] 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 is learned from sampling-based aggregates computed on the dataset without that individual.

Shoaran, Thomo, and Weber in [34], use ZKP to release connectedness statistics between groups in a social network. This is different from the current work, where we aim at privately releasing bridgeness statistics for linchpin nodes.

Regarding DP, [33] discusses the utility of the statistics distorted to satisfy DP. Here we consider the utility of the bridgeness statistics distorted to satisfy ZKP, and conclude that the utility is better than that of the ZKP mechanism in [34].

3 Graphs and Bridges

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 disjoint node groups of size r or more that a social network wants to release statistics about. Let g' and g'' be two groups in $\mathcal{S}$ and p be a node in G such that $p \notin g'$ and $p \notin g''$.

Definition 1. *The* bridgeness of node p on two groups g' and g'' *is defined as*

$$B_p(g', g'') = \frac{|\{(p, v', v'') : v' \in g', v'' \in g'', \text{ and } \{(v', v''), (p, v'), (p, v'')\} \subseteq E\}|}{|g'| \cdot |g''|}$$

Intuitively, bridgeness $B_p(g', g'')$ is the fraction of the number of (p, v', v'') triangles that exist over the number of all possible (p, v', v'') triangles.

Throughout the paper, we will refer to the bridgeness as B_p whenever g' and g'' are clear from the context.

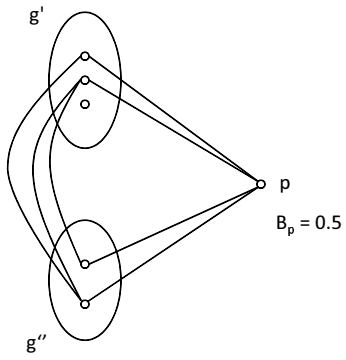

Fig. 1. Bridgeness

Example 1. Fig. 1 shows a graph G with two groups g' and g'', having three and two nodes, respectively. There are three edges connecting the nodes of g' and g'', and four edges connecting node p to the nodes of g' and g''. These edges form three triangles in total between p and two groups. The number of all possible such triangles is $3 \times 2 = 6$. Therefore, we have $B_p = \frac{3}{6} = 0.5$.

4 Background on ϵ-Zero-Knowledge Privacy

Zero-Knowledge Privacy (ZKP) introduced by [13] is an enhanced privacy scheme that guarantees stronger 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 settings 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 sanitized output of the mechanism is not more than what can be discovered by an assumed equally gifted agent that only has access to some sampling-based aggregate information. The latter agent is sometimes referred as *simulator*[4].

[4] In this context, adversaries and simulators are in fact some algorithms.

Thus, ZKP framework is defined based on a class of aggregate functions *agg*, such that the specification of those functions is used to define the privacy level of the ZKP mechanism. For example, the sample size in the class of aggregate functions directly affects the accuracy of the output (as it will be defined later in this section). Using such parameters we can design a ZKP mechanism that can provide a similar privacy protection. The importance of *agg* functions in the definition of ZKP is that by sampling data, the evidence of participation is also protected.

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 2. *(Zero-Knowledge Privacy* [13]*) 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:*

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

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 number of random samples – 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 [13]). 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, the set of groups $\mathcal{S}$, and a node p, f produces B_p measures for m pairs of groups. We consider the L_1-Sensitivity to be defined as follows.

Definition 3. *(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. *(Sample Complexity [13]) 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. Basically, 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 1. ([13]) *Suppose $f : \mathbf{G} \to [a, b]^m$ 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 Bridgeness

In this section we design a ZKP mechanism to privately release B_p measures. Let f be the function that given graph G, set $\mathcal{S}$, and node p produces a c-dimensional vector of B_p measures (numbers), where $c = \binom{|\mathcal{S}|}{2}$.

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 $(\sum_{i=1}^{t} \epsilon_i)$-ZKP with respect to $agg_{k(n)}$, where $k(n) = \sum_{i=1}^{t} k_i$.

Proposition 2. *(Sequential Composition [13]) 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*[5] *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. Therefore, we assume that the removed edge e is an edge between two nodes of two different groups in $\mathcal{S}$. Removing such an edge from G can change by at most 1 the numerator of a B_p measure in G_{-e}. Note that this change affects only one B_p measure in the whole graph G_{-e}. Therefore, the sensitivity of any B_p function is $\Delta(B_p) = 1/r^2$, where r is the minimum group size in $\mathcal{S}$.

Suppose $B_p(g, g')$ is an element of f, where g and g' are groups in $\mathcal{S}$. Let $San = B_p(g, g') + Lap(\lambda)$ be a ZKP mechanism which adds random noise selected from $Lap(\lambda)$ distribution to the output of $B_p(g, g')$ 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 B_p function. For this, without any change in semantics, we will express B_p so that it computes an average rather than a fraction of two counts. Then, using the *Hoeffding* inequality (cf. [26]) we compute the sample complexity of B_p.

Expressing B_p. In addition to regular node attributes (if any), we introduce $|\mathcal{S}|$ new boolean attributes, one for each group in $\mathcal{S}$. We denote each new attribute by upper-case I indexed by a group id. Each attribute I_g is a boolean vector of dimension $|g|$, where each dimension corresponds to a node in g. A node v in graph G will have $I_g(v)[u] = 1$, where $u \in g$, if $\{(v, u), (p, v), (p, u)\} \subseteq E$, and $I_g(v)[u] = 0$, otherwise. For each pair of groups g and g' we can show that

Proposition 3.

$$B_p(g, g') = \frac{\sum_{v \in g, u \in g'} I_{g'}(v)[u]}{|g| \cdot |g'|} = \frac{\sum_{v \in g', u \in g} I_g(v)[u]}{|g| \cdot |g'|}$$

Therefore, the $B_p(g, g')$ measure can be viewed as the average of $I_{g'}(v)[u]$'s or $I_g(v)[u]$'s.

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

[5] A set of computations that are separately applied on *one* database and each provides ZKP in isolation, also provides ZKP for the set.

$Lap(\lambda)$ distribution which will be used to add random noise to $B_p(g, g')$ measures included in f. For this, first we compute the sample complexity of B_p to be able to use Proposition 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[6] incident to these k sample nodes.[7] Node p is assumed to be included in the set of randomly selected nodes.

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

$$Pr[|T(g, g') - B_p(g, g')| \leq \delta] \geq 1 - 2e^{-2(k_g \times k_{g'})\delta^2}$$

From this and Definition 4, we have that B_p function has $\left(\delta, 2e^{-2K\delta^2}\right)$-sample complexity with respect to agg_k, where $K = (k_g \times k_{g'})$.

Now we make the following substitutions in the formula of Proposition 1: $\beta = 2e^{-2K\delta^2}$, $\Delta(B_p(g, g')) = 1/r^2$, $b - a = 1$, and $m = 1$. From this, we have that mechanism San is

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

with respect to agg_k.

Similarly to DP, we set λ, the Laplace noise scale, to be proportional to "the error" as can be measured in ZKP method by the sum of the sensitivity and sampling error, and inversely proportional to the ZKP privacy level.

$$\lambda = \frac{\Delta(B_p) + \delta}{\epsilon} = \frac{1}{\epsilon}\left(\frac{1}{r^2} + \frac{1}{\sqrt[3]{K}}\right)$$

Regarding δ, we can consider for instance a sample size $k(n) = \sqrt[3]{n^2}$, and have $\delta = \frac{1}{\sqrt[3]{K}}$.

From all the above, the privacy level obtained will be

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

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

[7] For other possible methods of graph sampling see for example [13].

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

Example 2. 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 the requested output vector is

$$f = \langle B_p(g', g''), B_p(g, g'')\rangle.$$

and suppose that the minimum group size in $\mathcal{S}$ is $r = 100$.

Assume we would like to have for f a ZKP mechanism expressed with respect to an acceptable agg_k, where

$$k(n) = \sqrt[3]{10,000,000^2} = 46,416.$$

To privately release the first output in f, a randomized algorithm T would uniformly select

$$k_1 = k(n)/2 = \sqrt[3]{10,000,000^2}/2 = 23,208.$$

nodes and approximate the value of $B_p(g', g'')$ using sample data.

The actual value of function $B_p(g', g'')$ is computed on G. Suppose that the size of the sample groups corresponding to g' and g'' are $k_{g'} = 500$ and $k_{g''} = 100$, respectively. Therefore, we have $K = 50,000$. Let (δ_1, β_1) be the sample complexity of $B_p(g', g'')$ where

$$\delta_1 = \frac{1}{\sqrt[3]{K}} = \frac{1}{\sqrt[3]{50,000}} = 0.0271.$$

$$\beta_1 = 2e^{-2K(\delta_1)^2} = 2e^{-2*(50,000)*(0.0271)^2} = 2.55 * 10^{-32}.$$

The sensitivity of f is

$$\Delta(f) = \frac{1}{r^2} = \frac{1}{100^2} = 0.0001.$$

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

$$\lambda_1 = \frac{\Delta(f) + \delta_1}{\epsilon} = \frac{0.0001 + 0.0271}{0.1} = 0.272$$

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

$$\epsilon_1 \leq \left(\epsilon + 2e^{-\sqrt[3]{K}}\right) = \left(0.1 + 2 * e^{-37}\right) \approx 0.1$$

with respect to agg_k.

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 $\delta = \frac{1}{\sqrt[3]{K}}$, where $K = k_{g'} * k_{g''}$.

Fig. 2 illustrates the relationship between the noise scale λ and the parameter 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 K (the product of group sample sizes) 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. Moreover, even $K = 500,000$ implies some sample group sizes, for example $k_{g'} = 1000$ and $k_{g''} = 500$, which are reasonable in social graphs with only millions of participants (see Example 2). Hence, we conclude that the proposed ZKP mechanism works well with small as well as large data graphs.

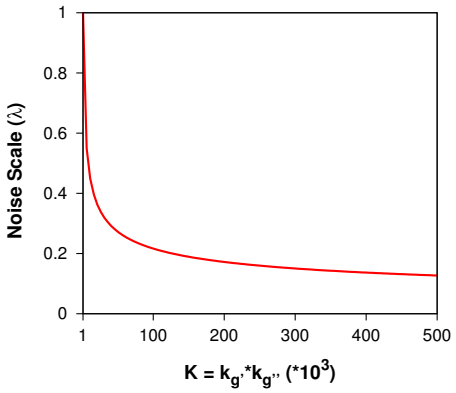

Fig. 2. Relationship between noise scale and sample group size.

6.2 The Noise

We present the analysis in this section in order 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,

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

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 3 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. These values of δ are practical as our outputs are fractions.

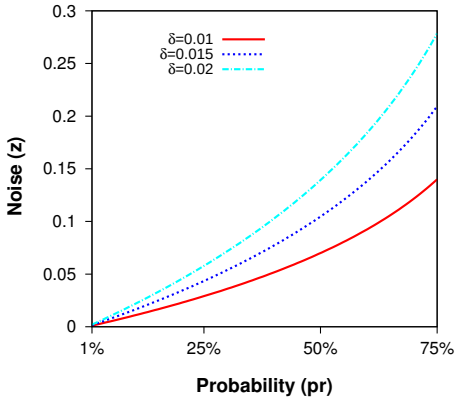

Fig. 3. Probability vs noise.

7 Conclusions

We addressed zero-knowledge privacy for releasing the bridgeness measure of graph nodes. The application of our technique is crucial in order to have a secure public release of graph properties. 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 data graphs. This is different from the mechanism presented in [34], which is useful only for

very large social graphs. As future work we aim at charting the landscape of other various graph statistics in order to determine their sample complexity and see for what sizes of graphs it makes sense to use ZKP from a utility point of view, i.e. without distorting too much the released statistics.

References

1. A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *PODS*, pages 128–138, 2005.
2. D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Y. Vardi. Reasoning on regular path queries. *SIGMOD Record*, 32(4):83–92, 2003.
3. D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Y. Vardi. View-based query processing: On the relationship between rewriting, answering and losslessness. *Theor. Comput. Sci.*, 371(3):169–182, 2007.
4. D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Y. Vardi. Query processing under glav mappings for relational and graph databases. *PVLDB*, 6(2):61–72, 2012.
5. X. Cheng, F. Ren, H. Shen, Z. Zhang, and T. Zhou. Bridgeness: A local indexon edge significance in maintaining global connectivity. *J. Stat. Mech.*, 10:10011, 2010.
6. S. Chester, B. M. Kapron, G. Ramesh, G. Srivastava, A. Thomo, and S. Venkatesh. k-anonymization of social networks by vertex addition. In *ADBIS*, pages 107–116, 2011.
7. S. Chester, B. M. Kapron, G. Ramesh, G. Srivastava, A. Thomo, and S. Venkatesh. Why waldo befriended the dummy? k-anonymization of social networks with pseudo-nodes. *SNAM*, 2012.
8. C. Dwork. Differential privacy. In *ICALP (2)*, pages 1–12, 2006.
9. C. Dwork. Differential privacy: A survey of results. In *TAMC*, pages 1–19, 2008.
10. C. Dwork. Differential privacy in new settings. In *SODA*, pages 174–183, 2010.
11. C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.
12. L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.
13. J. Gehrke, E. Lui, and R. Pass. Towards privacy for social networks: A zero-knowledge based definition of privacy. In *TCC*, pages 432–449, 2011.
14. G. Grahne and A. Thomo. Approximate reasoning in semistructured data. In *KRDB*, 2001.
15. G. Grahne, A. Thomo, and W. W. Wadge. Preferential regular path queries. *Fundam. Inform.*, 89(2-3):259–288, 2008.
16. N. Hassanlou, M. Shoaran, and A. Thomo. Probabilistic graph summarization. In *WAIM*, pages 545–556, 2013.
17. M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In *ICDM*, pages 169–178, 2009.
18. M. Khezrzadeh, A. Thomo, and W. W. Wadge. Harnessing the power of "favorites" lists for recommendation systems. In *RecSys*, pages 289–292, 2009.
19. D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD Conference*, pages 193–204, 2011.
20. D. Kifer and A. Machanavajjhala. A rigorous and customizable framework for privacy. In *PODS*, pages 77–88, 2012.
21. N. Koochakzadeh, A. Sarraf, K. Kianmehr, J. G. Rokne, and R. Alhajj. Netdriller: A powerful social network analysis tool. In *ICDM Workshops*, pages 1235–1238, 2011.

22. N. Korovaiko and A. Thomo. Trust prediction from user-item ratings. *Social Netw. Analys. Mining*, 3(3):749–759, 2013.
23. K. Liu and E. Terzi. Towards identity anonymization on graphs. In *SIGMOD Conference*, pages 93–106, 2008.
24. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. *L*-diversity: Privacy beyond *k*-anonymity. *TKDD*, 1(1), 2007.
25. G. Madey, V. Freeh, and R. Tynan. Modeling the free/open source software community: A quantitative investigation, 2005.
26. M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.
27. A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE Symposium on Security and Privacy*, pages 173–187, 2009.
28. T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó. Fuzzy communities and the concept of bridgeness in complex networks. *Phys. Rev. E*, 77:016107, 2008.
29. V. Rastogi, M. Hay, G. Miklau, and D. Suciu. Relationship privacy: output perturbation for queries with joins. In *PODS*, pages 107–116, 2009.
30. G. Sabidussi. The centrality of a graph. *Psychometrika*, 31(4):581–603, 1966.
31. M. Shoaran and A. Thomo. Fault-tolerant computation of distributed regular path queries. *Theor. Comput. Sci.*, 410(1):62–77, 2009.
32. M. Shoaran and A. Thomo. Certain answers and rewritings for local regular path queries on graph-structured data. In *IDEAS*, pages 186–192, 2010.
33. M. Shoaran, A. Thomo, and J. H. Weber. Differential privacy in practice. In *Secure Data Management*, pages 14–24, 2012.
34. M. Shoaran, A. Thomo, and J. H. Weber-Jahnke. Zero-knowledge private graph summarization. In *BigData Conference*, pages 597–605, 2013.
35. D. C. Stefanescu and A. Thomo. Enhanced regular path queries on semistructured databases. In *EDBT Workshops*, pages 700–711, 2006.
36. Y. Tian, R. A. Hankins, and J. M. Patel. Efficient aggregation for graph summarization. In *SIGMOD Conference*, pages 567–580, 2008.
37. N. Zhang, Y. Tian, and J. M. Patel. Discovery-driven graph summarization. In *ICDE*, pages 880–891, 2010.
38. P. Zhao, X. Li, D. Xin, and J. Han. Graph cube: on warehousing and olap multidimensional networks. In *SIGMOD Conference*, pages 853–864, 2011.
39. B. Zhou, J. Pei, and W.-S. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *SIGKDD Explorations*, 10(2):12–22, 2008.