Statistica Sinica Preprint No: SS-2019-0215	
Title	Directed Networks with a Differentially Private Bi-degree Sequence
Manuscript ID	SS-2019-0215
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0215
Complete List of Authors	Ting Yan
Corresponding Author	Ting Yan
E-mail	tingyanty@mail.ccnu.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Directed Networks with a Differentially Private Bi-degree Sequence

Ting Yan

Central China Normal University

Abstract: Although many approaches have been developed for releasing network data with a differential privacy guarantee, inference in many network models with differential privacy data is still unknown or has not been properly explored. In this paper, we propose to release the bi-degree sequences of directed networks using the Laplace mechanism and make inference in the p_0 model, which is an exponential random graph model with the bi-degree sequence as its exclusively sufficient statistic. We show that the estimator of the parameters without the so-called denoised process is asymptotically consistent and normally distributed. This is in sharp contrast to some known results that valid inference such as the existence and consistency of the estimator requires denoising. Along the way, a new phenomenon is revealed in which an additional variance factor appears in the asymptotic variance of the estimator to account for noise. An efficient algorithm is proposed for finding the closet point lying in the set of all graphical bi-degree sequences under the global L_1 optimization problem. Numerical study demonstrates our theoretical findings.

Key words and phrases: Asymptotic normality, Consistency, Differentially private, p_0 model, Synthetic graph.

1. Introduction

As more and more network data (of all kinds, but especially social ones) are collected and made publicly available, data privacy has become an important issue in network data analysis since they may contain sensitive information about individuals and their relation-

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

ships (e.g., sexual relationships, email exchanges). Directly publishing these sensitive data with anonymized or un-anonymized nodes could cause severe privacy problems or even lead to legal actions. For example, Netflix released the Netflix Prize data-set for public analysis in 2007, which contains anonymized network data about the viewing habits of its members. Two years later, Netflix was involved in a lawsuit with one of its members, who had been victimized by privacy invasions, done by applying some de-anonymization techniques to re-identify individual informations using public datasets [Task and Clifton (2012)]. Nevertheless, the benefit to analyze them is significant for addressing a variety of important issues including disease transmission, fraud detection, precision marketing, among many others.

To prevent the confidence information from being disclosed as well as ensure effective analysis, sensitive network data must be carefully treated before being made public. Although it is easy to attack under the anonymization technique by releasing an anonymized isomorphic network [e.g., Backstrom et al. (2011)], some refined anonymization techniques have been proposed; see for example Campan and Truta (2009); Narayanan and Shmatikov (2009); Zhou et al. (2008). These methods transform the original graph into a new graph by adding/removing edges or clustering of nodes into groups. However, they depend on an attacker's background knowledge and may fail to protect the private information. Dwork et al. (2006) developed a rigorous privacy standard–*differential privacy* for randomized data releasing mechanisms to achieve privacy protection. An algorithm satisfying differential privacy requires that the outputs should not be significantly different if the inputs are similar. Differential privacy provides strong guarantees of privacy without making any assumption about the background knowledge of attackers, and it has been widely used as a privacy standard to release network data in recent years [e.g., Hay et al. (2009); Lu and

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Miklau (2014); Task and Clifton (2012); Jorgensen et al. (2016)].

Although many differentially private algorithms have been developed to release network data or their aggregate network statistics safely [e.g., Jorgensen et al. (2016); Lu and Miklau (2014); Nguyen et al. (2016); Task and Clifton (2012)], statistical inference with noisy network data is still in its infancy. How to accurately estimate model parameters and analyze asymptotic properties of their estimators using noisy data in many network models is still unknown or has not been properly explored. There have been some recent developments in inferences with a differentially private degree sequence of undirected graphs. Hay et al. (2009) used the Laplace mechanism to release the degree partition and proposed an efficient algorithm to find the solution that minimizes the L_2 distance between all possible graphical degree partitions and the noisy degree partition. With this post-processing step, they obtained an accurate estimate of the degree distribution of a graph. Karwa and Slavković (2016) used a discrete Laplace mechanism to release the degree sequence. By using the techniques for proving the consistency of the maximum likelihood estimator in the β-model in Chatterjee et al. (2011) and those for obtaining its asymptotic normality in Yan and Xu (2013), Karwa and Slavković (2016) proved that a differentially private estimator of the parameter in the β-model is consistent and asymptotically normally distributed. Moreover, they constructed an efficient algorithm to denoise the differentially private degree sequence by solving an L_1 optimization problem. Day et al. (2016) proposed some approaches based on aggregation and cumulative histograms to publish the degree distribution under node differential privacy. Sealfon and Ullman (2019) proposed an efficient algorithm for estimating the parameter of an Erdös-Rényi graph under node differential privacy.

In this paper, we focus on inference by using the differentially private bi-sequences

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

of directed networks. As pointed by Hay et al. (2009), it may fail to protect privacy if we directly release the degree sequence since some graphs have unique degree sequences. In some other scenarios, the bi-degrees of nodes are themselves sensitive information. For instances, the out-degree of an individual in a sexually transmitted disease network reveals sensitive information such as how many people are infected by him/her. In this case, it is essential to limit the disclosure of the bi-degrees. In this paper, we propose to use the Laplace mechanism to release the bi-degree sequence and conduct inferences using the noisy bi-sequence. The main contributions are given as follows. First, we show that the estimator of the parameter in the p_0 model based on the moment equation in which the unobserved original bi-degree sequence is directly replaced by the noisy bi-sequence, is consistent and asymptotically normal without the denoised process. This is in sharp contrast to some existing results [e.g., Fienberg et al. (2010); Karwa and Slavković (2016)], in which ignoring the noisy process can lead to inconsistency and even nonexistent of consistent parameter estimates. The p_0 model is an exponential random graph model with the bi-degree sequence as its exclusively sufficient statistic. During our study, a new phenomenon is revealed in which an additional variance factor appears in the asymptotic variance of the estimator when the noise becomes large. To the best of our knowledge, this is the first time to discover this phenomenon in the noisy network data analysis. We further show that the differentially private estimator corresponding to the denoised bi-sequence is also consistent and asymptotically normal. Second, we propose an efficient algorithm to denoise the noisy bi-sequence, which finds the closest point lying in the set of all possible graphical bi-degree sequences under the global L_1 optimization problem. The denoised bi-sequence can be used to obtain an accurate estimate of the degree distribution of a directed graph. Along the way, it also outputs a synthetic directed graph that can be

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

used to infer the graph structure. We remark that the denoised step is generally needed for valid estimation of graph structures, since the noisy bi-sequence may not be graphical. Finally, we provide simulation studies as well as analysis of there real data sets to illustrate the theoretical results.

For the rest of the paper, we proceed as follows. In Section 2, we first introduce a necessary background on differential privacy. Then we present the estimation in the p_0 model using the differentially private bi-sequence. In Section 3, we present the consistency and asymptotic normality of the differentially private estimator. In Section 4, we denoise the noisy bi-sequence and present the asymptotic properties of the estimator corresponding to the denoised bi-sequence. In Section 5, we carry out simulation studies to evaluate the theoretical results and analyze three real network datasets. We make the summary and further discussion in Section 6. All proofs of theorems are relegated into online supplementary material.

2. Estimation from a differentially private bi-degree sequence

Let G_n be a simple directed graph on $n \geq 2$ nodes labeled by "1, . . . , n". Here, "simple" means that there are no multiple edges and no self-loops in G_n. Let $A = (a_{i,j})$ be the adjacency matrix of G_n, where $a_{i,j}$ is an indictor variable of the directed edge from head node i to tail node j. If there exists a directed edge from i to j, then $a_{i,j} = 1$; otherwise $a_{i,j} = 0$. Since G_n is loopless, we set $a_{i,i} = 0$ for convenience. Let $d_i^+ = \sum_{j \neq i} a_{i,j}$ be the out-degree of node i and $d^+ = (d_1^+, \ldots, d_n^+)^\top$ be the out-degree sequence of the graph G_n. Similarly, define $d_i^- = \sum_{j \neq i} a_{j,i}$ as the in-degree of node i and $d^- = (d_1^-, \ldots, d_n^-)^\top$ as the in-degree sequence. The pair $d = ((d^+)^\top, (d^-)^\top)^\top$ or $\{(d_1^+, d_1^-), \ldots, (d_n^+, d_n^-)\}$ are called the bi-degree sequence.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

In this section, we first give a brief introduction to differential privacy. Then we release the bi-degree sequence under edge differential privacy and estimate the degree parameter in the p_0 model.

2.1 Differential privacy

Consider an original database D containing a set of records of n individuals. We focus on mechanisms that take D as input and output a sanitized database $S = (S_1, \ldots, S_k)$ for public use. The size of S may not be the same as D. A randomized data releasing mechanism $Q(\cdot|D)$ defines a conditional probability distribution on outputs S given D. Let ϵ be a positive real number and $\mathcal{S}$ denote the sample space of Q. The data releasing mechanism Q is *ϵ-differentially private* if for any two neighboring databases D_1 and D_2 that differ on a single element (i.e., the data of one person), and all measurable subsets B of $\mathcal{S}$ [Dwork et al. (2006)],

$$Q(S \in B|D_1) \le e^{\epsilon} \times Q(S \in B|D_2).$$

The privacy parameter ϵ, which is publicly available, is chosen by the data curator administering the privacy policy, and controls the trade-off between privacy and utility. Smaller value of ϵ means more privacy protection.

Differential privacy requires that the distribution of the output is almost the same whether or not an individual's record appears in the database. We illustrate why it protects privacy with an example. Suppose a hospital wants to release some statistics on the medical records of their patients to the public. In response, a patient may wish to make his record omitted from the study due to a privacy concern that the published results will reveal some of his/her personal information. Differential privacy alleviates this concern because whether or not the patient participates in the study, the probability

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

of a possible output is almost the same. From a theoretical point, any test statistic has nearly no power for testing whether an individual's data is in the original database or not; see Wasserman and Zhou (2010) for a rigourous proof.

What is being protected in the differential privacy is precisely the difference between two neighboring databases. Within network data, depending on the definition of the graph neighbor, *differential privacy* is divided into *node differential privacy* [Kasiviswanathan et al. (2013)] and *edge differential privacy* [Nissim et al. (2007)]. Two graphs are called neighbors if they differ in exactly one edge, then *differential privacy* is *edge differential privacy*. Analogously, we can define *node differential privacy* by letting graphs be neighbors if one can be obtained from the other by removing a node and its adjacent edges. Edge differential privacy protects edges not to be detected, whereas node differential privacy protects nodes together with their adjacent edges, which is a stronger privacy policy. However, it may be infeasible to design algorithms that are both node differential privacy and have good utility. As an example, Hay et al. (2009) showed that estimating node degrees are highly inaccurate under node differential privacy due to that the global sensitive in Definition 2 is too large (in the worst case having an order n) such that the output is useless. Following Hay et al. (2009), we use edge differential privacy here.

Let $\delta(G, G')$ be the number of edges on which G and G' differ. The formal definition of edge differential privacy is as follows.

Definition 1 (Edge differential privacy)**.** Let $\epsilon > 0$ be a privacy parameter. A randomized mechanism $Q(\cdot|G)$ is ϵ-edge differentially private if

$$\sup_{G,G' \in \mathcal{G}, \delta(G,G')=1} \sup_{S \in \mathcal{S}} \frac{Q(S|G)}{Q(S|G')} \le e^{\epsilon},$$

where $\mathcal{G}$ is the set of all directed graphs of interest on n nodes and $\mathcal{S}$ is the set of all possible outputs.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Let $f: \mathcal{G} \to \mathbb{R}^k$ be a function. The global sensitivity [Dwork et al. (2006)] of the function f, denoted Δf, is defined below.

Definition 2. (Global Sensitivity). Let $f: \mathcal{G} \to \mathbb{R}^k$. The global sensitivity of f is defined as

$$\Delta(f) = \max_{\delta(G,G')=1} \|f(G) - f(G')\|_1$$

where $\|\cdot\|_1$ is the L_1 norm.

The global sensitivity measures the worst case difference between any two neighboring graphs. The magnitude of noises added in the differentially private algorithm Q crucially depends on the global sensitivity. If the outputs are the network statistics, then a simple algorithm to guarantee EDP is the Laplace Mechanism [e.g., Dwork et al. (2006)] that adds the Laplace noise proportional to the global sensitivity of f.

Lemma 1. *(Laplace Mechanism). Suppose $f: \mathcal{G} \to \mathbb{R}^k$ is a output function in $\mathcal{G}$. Let $e_1, \ldots, e_k$ be independent and identically distributed Laplace random variables with density function $e^{-|x|/\lambda}/\lambda$. Then the Laplace Mechanism outputs $f(G) + (e_1, \ldots, e_k)$ is ϵ-edge differentially private, where $\epsilon = -\Delta(f) \log \lambda$.*

When $f(G)$ is integer, one can use a discrete Laplace random variable as the noise as in Karwa and Slavković (2016), where it has the probability mass function:

$$\mathbb{P}(X = x) = \frac{1-\lambda}{1+\lambda}\lambda^{|x|}, \quad x \in \{0, \pm 1, \ldots\}, \lambda \in (0, 1).$$

Lemma 1 still holds if the continuous Laplace distribution is replaced by the discrete Laplace distribution.

One nice property of differential privacy is that any function of a differentially private mechanism is also differentially private.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Lemma 2 (Dwork et al. (2006); Wasserman and Zhou (2010))**.** *Let f be an output of an ϵ-differentially private mechanism and g be any function. Then $g(f(G))$ is also ϵ-differentially private.*

By Lemma 2, any post-processing done on an output of a differentially private mechanism is also differentially private.

2.2 The differentially private bi-degree sequence

We use the discrete Laplace mechanism in Lemma 1 to release the bi-degree sequence $d = (d^+, d^-)$ under edge differential privacy. Note that $f(G_n) = (d^+, d^-)$. If we add or remove a directed edge $i \to j$ in G_n, then the out-degree of the head node i and the in-degree of the tail node j increase or decrease 1 each. Therefore, the global sensitivity for the bi-degree sequence is 2. The released steps are in Algorithm 1, where a differentially private bi-sequence is returned.

Algorithm 1: Releasing d

Data: The bi-degree sequence d and privacy parameter ϵ_n

Result: The differentially private bi-sequence z

Let $d = (d^+, d^-)$ be the bi-degree sequence of G_n;

for $i = 1 \to n$ **do**

  Generate two independent e_i^+ and e_i^- from discrete Laplace with $\lambda_n = \exp(-\epsilon_n/2)$;

  Let $z_i^+ = d_i^+ + e_i^+$ and $z_i^- = d_i^- + e_i^-$

end

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

2.3 Estimation based on the p_0 model

To conduct statistical inferences from a noisy bi-sequence, we need to specify a model on the original bi-degree sequence. If no prior information is given, we can model d according to the maximum entropy principle [Wu (1997)]. It forces the probability distribution on graph G_n into the exponential family distribution with the bi-degree sequence as the sufficient statistic, which admits the maximum entropy when the expectation of a bi-degree sequence is given. Hereafter, we refer to this model as the p_0 model. The subscript "0" means a simpler model than the p_1 model that contains an additional reciprocity parameter [Holland and Leinhardt (1981)]. The p_0 model can be represented as:

$$\mathbb{P}(G_n) = \frac{1}{c(\alpha, \beta)} \exp(\sum_i \alpha_i d_i^+ + \sum_j \beta_j d_j^-), \tag{2.1}$$

where $c(\alpha, \beta)$ is a normalizing constant, $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$ and $\beta = (\beta_1, \ldots, \beta_n)^\top$. The outgoingness parameter α_i characterizes how attractive the node is and the incomingness parameter β_i illustrates the extent to which the node is attracted to others as in Holland and Leinhardt (1981). Although the p_0 model looks simple, it is still useful to applications where only the bi-degree sequence is used. First, it can be served as null models for hypothesis testing [e.g., Holland and Leinhardt (1981); Fienberg and Wasserman (1981); Zhang and Chen (2013)]. Second, it can be used to re-construct networks and make statistical inferences in a situation in which only the bi-degree sequence is available due to privacy consideration [e.g., Helleringer and Kohler (2007)]. Third, it can be used as a preliminary analysis for choosing suitable statistics for network configurations [e.g., Robins et al. (2009)].

Since an out-edge from node i pointing to j is the in-edge of j coming from i, it leads to that the sum of out-degrees is equal to the sum of in-degrees. If one transforms (α, β) to $(\alpha - c, \beta + c)$, the probability distribution in (2.1) does not change. For the sake of the

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

identification of model parameters, we set $\beta_n = 0$ as in Yan et al. (2016). The p_0 model can be formulated by an array of mutually independent Bernoulli random variables $a_{i,j}$, $1 \le i \ne j \le n$ with probabilities [Yan et al. (2016)]:

$$\mathbb{P}(a_{i,j} = 1) = \frac{e^{\alpha_i + \beta_j}}{1 + e^{\alpha_i + \beta_j}}.$$

The normalizing constant $c(\alpha, \beta)$ is $\sum_{i \ne j} \log(1 + e^{\alpha_i + \beta_j})$. We use the following equations to estimate the degree parameter:

$$\begin{aligned} z_i^+ &= \sum_{j \ne i} \frac{e^{\alpha_i + \beta_j}}{1+e^{\alpha_i + \beta_j}}, \quad i = 1, \ldots, n, \\ z_j^- &= \sum_{i \ne j} \frac{e^{\alpha_i + \beta_j}}{1+e^{\alpha_i + \beta_j}}, \quad j = 1, \ldots, n-1, \end{aligned} \tag{2.2}$$

where z is the differentially private bi-sequence of Algorithm 1. The fixed point iteration algorithm can be used to solve the above system of equations. Since the discrete Laplace distribution is symmetrical with mean zero, the above equations are also the moment equations. Let $\theta = (\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_{n-1})^\top$. The solution $\widehat{\theta}$ to the equations (2.2) is the differentially private estimator of θ according to Lemma 2, where $\widehat{\theta} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n, \hat{\beta}_1, \ldots, \hat{\beta}_{n-1})^\top$ and $\hat{\beta}_n = 0$.

3. Asymptotic properties of the estimator

In this section, we present the consistency and asymptotical normality of the differentially private estimator. For a subset $C \subset \mathbb{R}^n$, let C^0 and $\overline{C}$ denote the interior and closure of C, respectively. For a vector $x = (x_1, \ldots, x_n)^\top \in R^n$, denote by $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$, the ℓ_∞-norm of x. For an $n \times n$ matrix $J = (J_{i,j})$, let $\|J\|_\infty$ denote the matrix norm induced by the ℓ_∞-norm on vectors in $\mathbb{R}^n$, i.e.

$$\|J\|_\infty = \max_{x \ne 0} \frac{\|Jx\|_\infty}{\|x\|_\infty} = \max_{1 \le i \le n} \sum_{j=1}^n |J_{i,j}|.$$

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Since the privacy parameter ϵ_n is generally small, we assume that ϵ_n is bounded by a fixed constant. This will simplify some notations.

Since the number of parameters increases with the number of nodes, classical statistical theories can not be directly applied to obtain the asymptotic results of estimator. We use the Newton method developed in Yan et al. (2016) to establish the consistency. Here we are dealing with not only the high dimensional issue but also the errors carried by noises while Yan et al. (2016) only considered the high dimensional issue. The idea of the proof for the existence and consistency of $\widehat{\theta}$ can be briefly described as follows. Define a system of functions:

$$
\begin{aligned}
F_i(\theta) &= z_i^+ - \textstyle\sum_{k=1;k\neq i}^n \frac{e^{\alpha_i+\beta_k}}{1+e^{\alpha_i+\beta_k}}, \quad i=1,\ldots,n, \\
F_{n+j}(\theta) &= z_j^- - \textstyle\sum_{k=1;k\neq j}^n \frac{e^{\alpha_k+\beta_j}}{1+e^{\alpha_k+\beta_j}}, \quad j=1,\ldots,n, \\
F(\theta) &= (F_1(\theta),\ldots,F_{2n-1}(\theta))^\top.
\end{aligned} \tag{3.3}
$$

Notice that the solution to the equation $F(\theta)=0$ is precisely the estimator. We construct the Newton iterative sequence: $\theta^{(k+1)} = \theta^{(k)} - [F'(\theta^{(k)})]^{-1}F(\theta^{(k)})$. If the initial value is chosen as the true value θ^*, then it is left to bound the error between the initial point and the limiting point to show the consistency. This is done by establishing a geometric convergence of rate for the iterative sequence. The details are in online supplementary material. The existence and consistency of $\widehat{\theta}$ is stated blow.

Theorem 1. *Assume that $A\sim\mathbb{P}_{\theta^*}$, where $\mathbb{P}_{\theta^*}$ denotes the probability distribution (2.1) on A under the parameter θ^*. If $\epsilon_n^{-1}e^{12\|\theta^*\|_\infty} = o((n/\log n)^{1/2})$, then with probability approaching one as n goes to infinity, the estimator $\widehat{\theta}$ exists and satisfies*

$$
\|\widehat{\theta}-\theta^*\|_\infty = O_p\left(\frac{1}{\epsilon_n}\frac{(\log n)^{1/2}e^{6\|\theta^*\|_\infty}}{n^{1/2}}\right) = o_p(1).
$$

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Further, if $\widehat{\theta}$ *exists, it is unique.*

Remark 1. The condition $\epsilon_n^{-1} e^{12\|\theta^*\|_\infty} = o((n/\log n)^{1/2})$ in Theorem 1 to guarantee the consistency of the estimator, exhibits an interesting trade-off between the privacy parameter ϵ_n and $\|\theta^*\|_\infty$. If $\|\theta^*\|_\infty$ is bounded by a constant, ϵ_n can be as small as $n^{1/2}/(\log n)^{-1/2}$. Conversely, if $e^{\|\theta^*\|_\infty}$ grows at a rate of $n^{1/12}/(\log n)^{1/12}$, then ϵ_n can only be at a constant magnitude.

In order to present asymptotic normality of $\widehat{\theta}$, we introduce a class of matrices. Given two positive numbers m and M with $M \geq m > 0$, we say the $(2n-1) \times (2n-1)$ matrix $V = (v_{i,j})$ belongs to the class $\mathcal{L}_n(m, M)$ if the following holds:

$$
\begin{aligned}
& m \leq v_{i,i} - \textstyle\sum_{j=n+1}^{2n-1} v_{i,j} \leq M, \quad i = 1, \ldots, n-1; \quad v_{n,n} = \textstyle\sum_{j=n+1}^{2n-1} v_{n,j}, \\
& v_{i,j} = 0, \quad i, j = 1, \ldots, n, \ i \neq j, \\
& v_{i,j} = 0, \quad i, j = n+1, \ldots, 2n-1, \ i \neq j, \\
& m \leq v_{i,j} = v_{j,i} \leq M, \quad i = 1, \ldots, n, \ j = n+1, \ldots, 2n-1, \ j \neq n+i, \\
& v_{i,n+i} = v_{n+i,i} = 0, \quad i = 1, \ldots, n-1, \\
& v_{i,i} = \textstyle\sum_{k=1}^{n} v_{k,i} = \textstyle\sum_{k=1}^{n} v_{i,k}, \quad i = n+1, \ldots, 2n-1.
\end{aligned} \tag{3.4}
$$

Clearly, if $V \in \mathcal{L}_n(m, M)$, then V is a $(2n-1) \times (2n-1)$ diagonally dominant, symmetric nonnegative matrix. Define $v_{2n,i} = v_{i,2n} := v_{i,i} - \sum_{j=1; j \neq i}^{2n-1} v_{i,j}$ for $i = 1, \ldots, 2n-1$ and $v_{2n,2n} = \sum_{i=1}^{2n-1} v_{2n,i}$. Yan et al. (2016) propose to approximate the inverse of V, V^{-1}, by

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

the matrix $S = (s_{i,j})$, which is defined as

$$s_{i,j} = \begin{cases} \frac{\delta_{i,j}}{v_{i,i}} + \frac{1}{v_{2n,2n}}, & i,j = 1,\ldots,n, \\ -\frac{1}{v_{2n,2n}}, & i = 1,\ldots,n, \quad j = n+1,\ldots,2n-1, \\ -\frac{1}{v_{2n,2n}}, & i = n+1,\ldots,2n-1, \quad j = 1,\ldots,n, \\ \frac{\delta_{i,j}}{v_{i,i}} + \frac{1}{v_{2n,2n}}, & i,j = n+1,\ldots,2n-1, \end{cases} \tag{3.5}$$

where $\delta_{i,j} = 1$ when $i = j$ and $\delta_{i,j} = 0$ when $i \neq j$.

We use V to denote the Fisher information matrix of θ in the p_0 model. It can be shown that

$$v_{ij} = \frac{e^{\alpha_i+\beta_j}}{(1+e^{\alpha_i+\beta_j})^2}, \quad 1 \le i \neq j \le n.$$

Since $e^x/(1+e^x)^2$ is an increasing function on x when $x \ge 0$ and a decreasing function when $x \le 0$, we have

$$\frac{(n-1)e^{2\|\theta\|_\infty}}{(1+e^{2\|\theta\|_\infty})^2} \le v_{ii} \le \frac{n-1}{4}, \quad i = 1,\ldots,2n.$$

Therefore $V \in \mathcal{L}_n(m, M)$, where m is the left expression and M is the right expression in the above inequality. The asymptotic distribution of $\widehat{\theta}$ depends on V. Let $g = (d_1^+,\ldots,d_n^+,d_1^-,\ldots,d_{n-1}^-)^\top$ and $\tilde{g} = (z_1^+,\ldots,z_n^+,z_1^-,\ldots,z_{n-1}^-)^\top$. If we apply Taylor's expansion to each component of $\tilde{g} - \mathbb{E}g$, then the second order term in the expansion is $V(\widehat{\theta} - \theta)$. Since V^{-1} does not have a closed form, we work with S defined at (3.5) to approximate it. Then we represent $\widehat{\theta} - \theta$ as the sum of $S(\tilde{g} - \mathbb{E}g)$ and a remainder. The central limit theorem is proved by establishing the asymptotic normality of $S(\tilde{g} - \mathbb{E}g)$ and showing the remainder is negligible. We formally state the central limit theorem as follows.

Theorem 2. *Assume that* $A \sim \mathbb{P}_{\theta^*}$ *and* $\epsilon_n^{-2} e^{18\|\theta^*\|_\infty} = o((n/\log n)^{1/2})$.

(i) If $\epsilon_n^{-1}(\log n)^{1/2} e^{2\|\theta^*\|_\infty} = o(1)$*, then for any fixed* $k \ge 1$*, as* $n \to \infty$*, the vector consisting*

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

of the first k elements of $(\widehat{\theta} - \theta^*)$ is asymptotically multivariate normal with mean $\mathbf{0}$ and covariance matrix given by the upper left $k \times k$ block of S defined at (3.5).

(ii) Let

$$s_n^2 = \mathrm{Var}(\sum_{i=1}^{n} e_i^+ - \sum_{i=1}^{n-1} e_i^-) = (2n-1)\frac{2e^{-\epsilon_n/2}}{(1-e^{-\epsilon_n/2})^2}.$$

If $s_n/v_{2n,2n}^{1/2} \to c$ for some constant c, then for any fixed $k \geq 1$, the vector consisting of the first k elements of $(\widehat{\theta} - \theta^*)$ is asymptotically k-dimensional multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\mathrm{diag}(\frac{1}{v_{1,1}}, \ldots, \frac{1}{v_{k,k}}) + (\frac{1}{v_{2n,2n}} + \frac{s_n^2}{v_{2n,2n}^2})\mathbf{1}_k\mathbf{1}_k^\top,$$

where $\mathbf{1}_k$ is a k-dimensional column vector with all entries 1.

Remark 2. First, if we change the first k elements of $(\widehat{\theta} - \theta^*)$ to an arbitrarily fixed k elements with the subscript set $\{i_1, \ldots, i_k\}$, Theorem 2 still holds. This is because all steps in the proof are valid if we change the first k subscript set $\{1, \ldots, k\}$ to $\{i_1, \ldots, i_k\}$. Second, the asymptotic variance for the difference of the pairwise estimators $(\widehat{\theta} - \theta^*)_i - (\widehat{\theta} - \theta^*)_j$ is $1/v_{i,i} + 1/v_{j,j}$, regardless of the additional variance factor $1/v_{2n,2n} + s_n^2/v_{2n,2n}^2$.

Remark 3. In the second part of Theorem 2, the asymptotic variance of $\widehat{\theta}_i$ has an additional variance factor $s_n^2/v_{2n,2n}^2$. This is different from Theorem 2 in Yan et al. (2016), in which they consider the a non-differential private case. The asymptotic expression of $\hat{\theta}_i$ contains a term $\sum_{i=1}^n e_i^+ - \sum_{i=1}^{n-1} e_i^-$. Its variance is in the magnitude of $ne^{-\epsilon_n/2}$. When ϵ_n becomes small, the variance increases quickly such that its impact on the $\widehat{\theta}_i$ can not be ignored when it increases to a certain level. This leads to the appearance of the additional variance factor.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

4. The denoised bi-degrees and synthetic directed graphs

The output z of Algorithm 1 generally is not the graphical bi-degree sequence. There have been several characterizations for the bi-degree sequence [e.g., Fulkerson (1960); Kleitman and Wang (1973); Majcher (1985)]. A necessary condition for graphical bi-degree sequences is that the sum of in-degrees is equal to that of out-degrees and all in- and out- degrees are between 0 and $n-1$. To check what are the chances that this condition holds, we carry out some simulations. We use the p_0 model to generate the random graphs and record their bi-degree sequences. Then use Algorithm 1 to output the bi-sequence z. We set $\alpha_i, \beta_i \sim U(0,1)$ and $n=100$. We repeat $10,000$ simulations and record the frequency that $\sum_i z_i^+ = \sum_i z_i^-$ holds. The simulation results show that this condition holds with at most 1%.

To make z be graphical, we need to denoise z. The denoising process appears to be complex. First, the number of parameters to be estimated $(d_i^+, d_i^-, i=1,\ldots,n)$ is equal to the number of observations $(z_i^+, z_i^-, i=1,\ldots,n)$. Second, the parameter space is discrete and very large, whose cardinality grows at least an exponential magnitude. Let B_n be the set of all possible bi-degree sequence of graph G_n. It is natural to use the closest point $\hat{d}$ lying in B_n as the denoised bi-sequence with some distance between $\hat{d}$ and d. We use L_1 distance here and define the estimator as

$$\hat{d} = \arg\min_{d\in B_n} (\|z^+ - d^+\|_1 + \|z^- - d^-\|_1). \tag{4.6}$$

Notice that the maximum likelihood estimation leads to the same solution. Specifically, since the parameter λ_n in the noise addition process of Algorithm 1 is known, the likelihood on observation z with the parameter d in B_n is

$$L(d|z) = c(\lambda_n)\exp\{-(\sum_{i=1}^{n} |z_i^+ - d_i^+| + \sum_{i=1}^{n-1} |z_i^- - d_i^-|)\}.$$

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

We can see that the MLE of d is also $\hat{d}$.

We propose Algorithm 2 to produce the MLE $\hat{d}$. Along the way, it also outputs a directed graph with $\hat{d}$ as its bi-degree sequence. The correctness of Algorithm 2 is given in Theorem 3, whose proof is in online supplementary material.

Algorithm 2: Denoising z

Data: A bi-sequence of integers $z = (z^+, z^-)$

Result: A directed graph G_n on n vertices with bi-degree sequence $\hat{d}$

Let G_n be the empty graph on n vertices;

Let $S = \{1, \ldots, n\} \setminus \{i : z_i^+ \leq 0\}$;

while $|S| > 0$ **do**

  $T = \{1, \ldots, n\} \setminus \{i : z_i^- \leq 0\}$;

  Let $z_{i^*}^+ = \max_{i \in S} z_i^+$ and $i^* = \min\{i \in S : z_i^+ = z_{i^*}^+\}$;

  Let $T = T \setminus \{i^*\}$ and $pos = |T|$;

  Let $h_{i^*} = \min(z_{i^*}^+, pos)$;

  Let I =indices of h_{i^*} highest values in $z^-(T)$ where $z^-(T)$ is the sequence z^-;

  restricted to the index set T;

  Add a directed edge from i^* to k in G_n for each $k \in I$;

  Let $z_i^- = z_i^- - 1$ for all $i \in I$ and $S = S \setminus \{i^*\}$

end

Theorem 3. *Let $z = (z^+, z^-)$ be a bi-sequence of integers obtained from Algorithm 1. The bi-degree sequence of G_n produced by Algorithm 2 is $\hat{d}$ defined at (4.6).*

We prove Theorem 3 by converting the directed Havel-Hakimi algorithm [Erdős et al. (2010)] into Algorithm 2 that performs L_1 "projection" on the set B_n. This is motivated by

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Karwa and Slavković (2016) who used the Havel-Hakimi algorithm [Havel (1955); Hakimi (1962)] to find the solution to the undirected L_1 optimalization problem. Although the Havel-Hakimi algorithm had been proposed sixty years ago, the directed version has been derived until Erdós et al. (2010). In the directed case, one needs to consider the in-degree sequence and out-degree sequence simultaneously. Therefore, our algorithm is not a trivial extension of the algorithm in the undirected case in Karwa and Slavković (2016).

Remark 4. In step 8 of Algorithm 2, if some in-degrees of $z^-(T)$ are equal, we arrange them by the decreasing order of their corresponding out-degrees. Assume that the order is $z^-_{i_1} \geq \cdots \geq z^-_{i_k}$. Then we select their top h_{i^*} values. This rule applies hereafter and we will not emphasize it.

The next theorem characterizes the error between $\hat{d}$ and d in terms of the privacy parameter ϵ_n.

Theorem 4. *When $\epsilon_n(c+1) \geq 4\log n$, we have*

$$\mathbb{P}(\|\hat{d} - d\|_\infty > c) \leq \frac{4}{n},$$

where for two bi-sequences $a = (a^+, a^-)$ and $b = (b^+, b^-)$, $\|a - b\|_\infty$ is defined as

$$\|a - b\|_\infty = \max\{\|a^+ - b^+\|_\infty, \|a^- - b^-\|_\infty\} \tag{4.7}$$

As expected, the smaller the privacy parameter ϵ_n is, the larger the error between the original bi-degree and its MLE $\hat{d}$ will be. For any fixed $\tau \in (0, 1/2)$, if $\epsilon_n = \Omega(n^{-(1/2-\tau)})$, then

$$\|\hat{d} - d\|_\infty = O_p(n^{(1/2-\tau)} \log n). \tag{4.8}$$

Both $\tilde{d}$ and $\hat{d}$ are the EDP estimator of d, where the latter is due to Lemma 2. We can use $\hat{d}$ to replace $\tilde{d}$ in equations (2.2) to obtain the denoised estimator of the parameter

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

θ and denote the solution as $\bar{\theta}$. By repeatedly using Lemma 2, $\widehat{\theta}$ and $\bar{\theta}$ are both EDP estimators. By noting (4.8) holds, with the similar lines of arguments for Theorems 1 and 2, $\bar{\theta}$ is also consistent and asymptotically normal stated in Theorem 5, whose proof is given in the supplementary material.

Theorem 5. *Assume that $A \sim \mathbb{P}_{\theta^*}$.*

(i) If $e^{12\|\theta^*\|_\infty} = o((n/\log n)^{1/2})$ and $\epsilon_n = \Omega((\log n/n)^{1/2})$, then as n goes to infinity, with probability approaching one, the EDP estimator $\bar{\theta}$ exists and satisfies

$$\|\bar{\theta} - \theta^*\|_\infty = O_p\left(\frac{(\log n)^{1/2} e^{6\|\theta^*\|_\infty}}{n^{1/2}}\right) = o_p(1).$$

Further, if $\bar{\theta}$ exists, it is unique.

(ii) If $e^{18\|\theta^*\|_\infty} = o((n/\log n)^{1/2})$ and $\epsilon_n^{-1} e^{6\|\theta^*\|_\infty} = o(n^{1/2}/\log n)$, then for any fixed $k \geq 1$, as $n \to \infty$, the vector consisting of the first k elements of $(\bar{\theta} - \theta^*)$ is asymptotically multivariate normal with mean $\mathbf{0}$ and covariance matrix given by the upper left $k \times k$ block of S defined at (3.5).

Remark 5. Since the distribution of the difference $\hat{d} - d$ is difficult to obtain, we don't have the asymptotic result like in Theorem 2 (ii). By Theorem 5, the convergence rate of $\bar{\theta}_i$ is $1/v_{i,i}^{1/2}$ for any fixed i. Since $(n-1)e^{-2\|\theta^*\|_\infty}/4 \leq v_{i,i} \leq (n-1)/4$, the rate of convergence is between $O(n^{-1/2}e^{\|\theta^*\|_\infty})$ and $O(n^{-1/2})$, which is the same as the non private estimator [Yan et al. (2016)].

5. Numerical studies

The simulation results to assess the performance of the estimator for finite sizes of networks under different n, ϵ_n and θ are given in online supplementary material. Three real data analyses are also provided. We only present one real data analysis here and the other

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

two are put in the supplementary material.

5.1 Real data analysis

We evaluate how close the estimator $(\hat{\alpha}, \hat{\beta})$ is to the MLE $(\tilde{\alpha}, \tilde{\beta})$ fitted in the p_0 model with the original bi-degree sequence through three real network datasets, which are the Children's Friendship data, Lazega's Law Firm data and Uc irvine messages data, respectively. We only present the analytical results of the Uc irvine messages data here and the others are put in supplementary material. Note that $(\hat{\alpha}, \hat{\beta})$ is the edge differentially private estimator of the vector parameters α and β. If only the private estimator is released, then whether an edge is present or not in the original dataset could almost not be detected. We chose ϵ_n equal to 1, 2 and 3 as in Karwa and Slavković (2016) and repeated to release the bi-degree sequence using Algorithm 1 $1,000$ times for each ϵ_n. Then we computed the average private estimate and the upper (97.5^{th}) in blue color and the lower (2.5^{th}) quantiles in orange color of the estimates conditional on the event that the private estimate exists.

The Uc irvine messages network data was collected from an online community of students at the University of California, Irvine [Opsahl and Panzarasa (2009)]. It has a total of 1899 nodes and each node represents a student. A directed edge is established from one student to another if one or more messages have been sent from the former to the latter. A total of $20,296$ edges form and the edge density is 0.56%, indicating a very sparse network. Among $1,899$ nodes, there are 586 nodes having no out-edges or in-edges. We remove them due to that the non private MLE does not exist in this case. To guarantee non zero out-degrees and in-degrees after adding noises with a large probability, we only analyze a subgraph with their out-degrees and in-degrees both larger than 5. After data

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

preprocessing, only 696 nodes are left and the quantiles of 0, 1/4, 1/2, 3/4, 1 are 3, 8, 14, 26, 164 for out-degrees and 4, 10, 16, 27, 121 for in-degrees, respectively.

When many nodes have few links to others, large noise is easy to cause the output with non positive elements in Algorithm 1. When $\epsilon = 1$, the average ℓ_∞-distance between d and $\tilde{d}$ is 15.6 and all private estimates fail to exist. In this case, we try another $\epsilon = \log n/n^{1/4}$ (≈ 1.27). The frequencies that the private estimate fails to exist are 99.3%, 54.9% and 8.3% for $\epsilon = \log n/n^{1/4}, 2, 3$, respectively. The results are shown in Figure 1. From this figure, we can see that the mean value of $\hat{\alpha}$ or $\hat{\beta}$ are very close to the MLE and the MLE still lies in the 95% confidence interval.

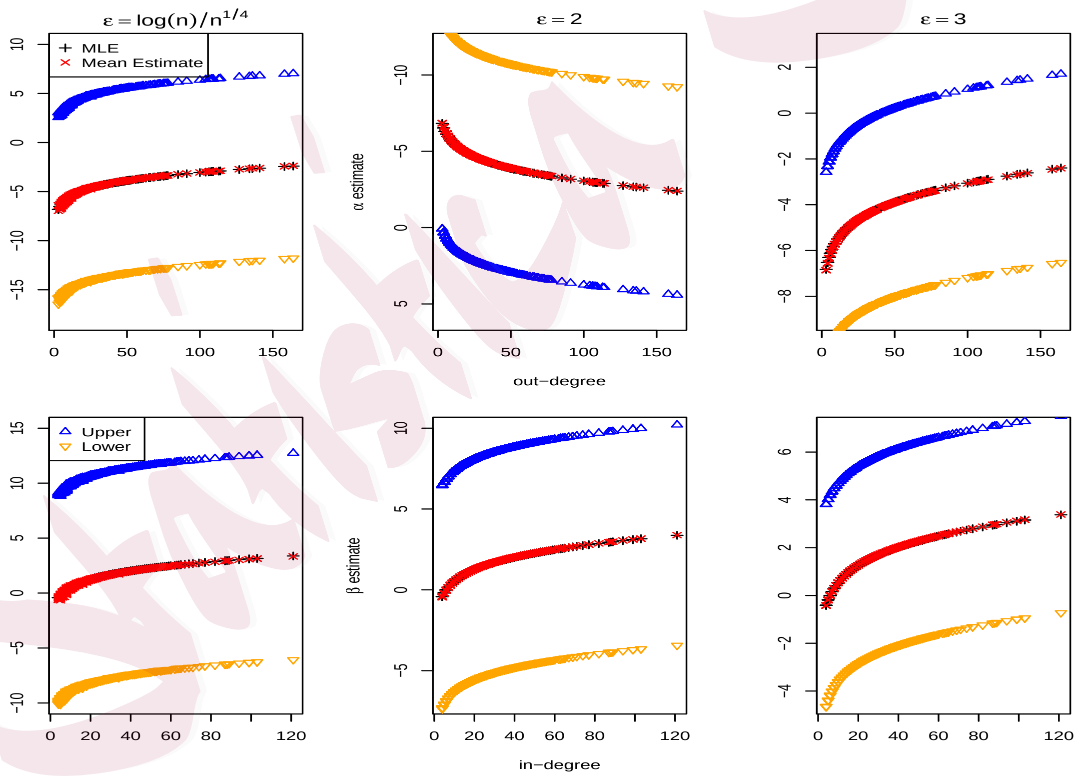

Figure 1: The differentially private estimate $(\hat{\alpha}, \hat{\beta})$ with the MLE for the Uc irvine messages network.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

6. Discussion

We have presented the consistency of the differentially private estimator of the parameter in the p_0 model under some mild conditions when the discrete Laplace noise is added into the bi-degree. We have revealed a phase transition for the asymptotic variance of the estimator in which an additional variance factor appears when the variance of the noise increases. The simulation shows that ignoring it could lead to invalid conference intervals. The added noise introduces considerable error when applying the noisy bi-sequence to estimate the degree distribution. We propose an efficient algorithm to denoise the noisy bi-sequence. The denoised bi-sequence can be used to obtain an accurate estimate of the degree distribution of a directed graph. Our simulation studies show that the non denoised estimator has a better performance than the denoised estimator for finite network sizes. On the other hand, when the privacy parameter ϵ_n is small, the private estimate fails to exist with positive frequencies according to our numerical studies, especially when the network dataset is sparse. An approach to avoid this problem is adding positive Laplace random noises or using f-differential privacy. We would like to investigate this problem in the future.

The conditions in Theorems 1 and 2 induce an interesting trade-off between the private parameter measuring the magnitude of the noise and the growing rate of the parameter θ. If the parameter ϵ_n is large, θ can be allowed to be relatively large. For instance, if $\epsilon_n = O(1)$, then the condition (i.e., $(1 + 4\epsilon_n^{-1})e^{12\|\theta^*\|_\infty} = o((n/\log n)^{1/2})$) in Theorem 1 becomes $e^{12\|\theta^*\|_\infty} = o((n/\log n)^{1/2})$. Moreover, the condition in Theorem 2 is much stronger than that in Theorem 1. The asymptotic behavior of the estimator is not only determined by the growing rate of the parameter θ, but also by the configuration of the parameter. It would be of interest to see whether these conditions can be relaxed.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

There are two different tasks for data privacy problem. The first is data protection. If the network model contains other network features such as k-stars and triangle and only these network statistics are of interest, then the additive noisy mechanism in this paper can be used to disclose them safely and it satisfied the edge differential privacy if the Laplace noise is added. The second is making inference from the noisy data. In order to extend the method of deriving the consistency of the estimator in our paper to other network models, one needs to establish a geometrical rate of convergence of the Newton iterative sequence. This is not easy for network models with other network features since it is difficult to derive the upper bound of the matrix norm for the inverse matrix of the Fisher information matrix without some special matrix structures. At the same time, it is also difficult to extend the method of deriving asymptotic normality of the estimator to network models with other network features since it is generally difficult to derive the approximate inverse matrix of a general Fisher information matrix.

Supplementary Materials

The online supplementary materials contain the simulation results, two real data analyses and the proofs of Theorems 1–5.

Acknowledgements

I am very grateful to two anonymous referees, an associated editor and the Editor for their valuable comments that have greatly improved the manuscript. I thank Binyan Jiang in Hong Kong Polytechnic University and Chenlei Leng in University of Warwick for carefully reading the manuscript and providing helpful comments. This work is partially supported by the National Natural Science Foundation of China (No. 11771171) and the Fundamental Research Funds for the Central Universities (No. CCNU17TS0005).

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

References

Backstrom L., Dwork C. and Kleinberg J. (2011). Wherefore art thou R3579X?: anonymized social networks, hidden patterns, and structural steganography. *Commun. ACM*, 54, 133-141.

Chatterjee S., Diaconis P., and Sly A. (2011). Random graphs with a given degree sequence. *Annals of Applied Probability*, **21**, 1400–1435.

Campan A. and Truta T. M. (2009). Data and Structural k-Anonymity in Social Networks. *Privacy, Security, and Trust in KDD* edited by Bonchi, Francesco and Ferrari, Elena and Jiang, Wei and Malin, Bradley. Springer Berlin Heidelberg, Berlin, Heidelberg, 33–54.

Day W., Li N. and Lyu M. (2016). Publishing graph degree distribution with node differential privacy. *In Proceedings of the 2016 International Conference on Management of Data*, 123–138, ACM, NY, USA.

Dwork C., Mcsherry F., Nissim K. and Smith A. (2006). Calibrating noise to sensitivity in private data analysis. *Proceedings of the 3rd Theory of Cryptography Conference*, 265–284.

Erdós P. L., Péter L. Miklós I., and Toroczkai, Z. (2010) A simple Havel-Hakimi type algorithm to realize graphical degree sequences of directed graphs. *The Electronic Journal of Combinatorics*, 17, Research Paper R66.

Fienberg S. E., Rinaldo A. and Yang X. (2010). Differential privacy and the risk- utility tradeoff for multi-dimensional contingency tables. In *Proceedings of the 2010 International Conference on Privacy in Statistical Databases*, PSD'10 187-199. Springer, Berlin.

Fienberg, S. E. and Wasserman, S. (1981). An exponential family of probability distributions for directed graphs: comment. *Journal of the American Statistical Association*, **76**(373), 54–57.

Fulkerson D. R. (1960). Zero-one matrices with zero trace. *Pacific J. Math.*, 10, 831–836.

Hakimi S. L. (1962). On realizability of a set of integers as degrees of the vertices of a linear graph. I. *Journal of the Society for Industrial and Applied Mathematics*, 496–506.

Havel V. (1955). A remark on the existence of finite graphs. *Časopis pro pěstování matematiky*, 80, 477–480.

Hay M., Li C., Miklau G. and Jensen D. (2009). Accurate estimation of the degree distribution of private networks. In Data Mining, 2009. ICDM09. Ninth IEEE International Conference on 169C178. IEEE.

Helleringer S, Kohler HP. (2007). Sexual network structure and the spread of HIV in

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Africa: evidence from Likoma Island, Malawi. *AIDS* 2007,21(17):2323–32.

Holland P. W. and Leinhardt S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, **76**, 33–65.

Jorgensen Z., Yu T. and Cormode G. (2016). Publishing Attributed Social Graphs with Formal Privacy Guarantees. *Proceedings of the 2016 International Conference on Management of Data*, 107–122. ACM, NY, USA.

Kasiviswanathan S.P., Nissim K., Raskhodnikova S., Smith A. (2013). Analyzing Graphs with Node Differential Privacy. In: Sahai A. (eds) Theory of Cryptography. Lecture Notes in Computer Science, vol 7785. Springer, Berlin, Heidelberg.

Karwa V. and Slavković A. (2016). Inference using noisy degrees-Differentially private beta model and synthetic graphs. *The Annals of Statistics*, **44**, 87–112.

Kleitman D. and Wang D. (1973). Algorithms for constructing graphs and digraphs with given valences and factors. *Discrete Mathmatics*, 6, 79–88.

Lu W. and Miklau G. (2014). Exponential random graph estimation under differential privacy. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*, ACM, New York, NY, USA, 921–930.

Majcher Z. (1985). Matrices representable by directed graphs. *Archivum Mathematicum*, 4, 205–218.

McCormick T. H., Salganik M. J. and Zheng T. (2010). How Many People Do You Know?: Efficiently Estimating Personal Network Size. *Journal of the American Statistical Association*, 105, 59–70.

Narayanan A. and Shmatikov V. (2009). De-anonymizing Social Networks. *30th IEEE Symposium on Security and Privacy*, Berkeley, CA, pp. 173-187.

Nguyen H., Imine A. and Rusinowitch M. (2016). Detecting communities under differential privacy. *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*, 83–93. ACM, NY, USA.

Nissim K., Raskhodnikova S. and Smith A. (2007). Smooth sensitivity and sampling in private data analysis. *In Proceedings of the thirty-ninth annual ACM Symposium on Theory of Computing*, 75–84. ACM.

Day W., Li N. and Lyu M. (2016). Publishing graph degree distribution with node differential privacy. *In Proceedings of the 2016 International Conference on Management of Data*, 123–138, ACM, NY, USA.

Opsahl T. and Panzarasa P. (2009). Clustering in weighted networks. *Social Networks*, 31, 155–163.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Robins G., Pattison P., and Wang P. (2009). Closure, connectivity and degree distributions: Exponential random graph (p^*) models for directed social networks. *Social Networks*, **31**, 105–117.

Sealfon A. and Ullman J. (2019). Efficiently estimating erdös-rényi graphs with node differential privacy. Available at arXiv:1905.10477

Task C. and Clifton C. (2012). A guide to differential privacy theory in social network analysis. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 411–417.

Wasserman L. and Zhou S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association* **105**, 375–389.

Wu N. (1997). The maximum entropy method. New York, Springer.

Yan T., Leng C. and Zhu J. (2016). Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. *The Annals of Statistics*, **44**, 31–57.

Yan T. and Xu J. (2013). A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices. *Biometrika*, **100**, 519–524.

Zhang J. and Chen Y. (2013). Sampling for conditional inference on network data. *Journal of the American Statistical Association*, 108, 1295–1307.

Zhou B., Pei J. and Luk W. (2008). A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter archive*, 10, 12–22.

Ting Yan
Department of Statistics
Central China Normal University
Wuhan, 430079
China
E-mail: tingyanty@mail.ccnu.edu.cn