

INFERENCE USING NOISY DEGREES: DIFFERENTIALLY PRIVATE β -MODEL AND SYNTHETIC GRAPHS

BY VISHESH KARWA^{1,2} AND ALEKSANDRA SLAVKOVIĆ¹

Carnegie Mellon University and Pennsylvania State University

The β -model of random graphs is an exponential family model with the degree sequence as a sufficient statistic. In this paper, we contribute three key results. First, we characterize conditions that lead to a quadratic time algorithm to check for the existence of MLE of the β -model, and show that the MLE never exists for the degree partition β -model. Second, motivated by privacy problems with network data, we derive a differentially private estimator of the parameters of β -model, and show it is consistent and asymptotically normally distributed—it achieves the same rate of convergence as the nonprivate estimator. We present an efficient algorithm for the private estimator that can be used to release synthetic graphs. Our techniques can also be used to release degree distributions and degree partitions accurately and privately, and to perform inference from noisy degrees arising from contexts other than privacy. We evaluate the proposed estimator on real graphs and compare it with a current algorithm for releasing degree distributions and find that it does significantly better. Finally, our paper addresses shortcomings of current approaches to a fundamental problem of how to perform valid statistical inference from data released by privacy mechanisms, and lays a foundational groundwork on how to achieve optimal and private statistical inference in a principled manner by modeling the privacy mechanism; these principles should be applicable to a class of models beyond the β -model.

Received August 2014; revised June 2015.

¹Supported in part by NSF Grant BCS-0941553 to the Department of Statistics, Pennsylvania State University.

²Supported in part by the Singapore National Research Foundation under its International Research Centre Singapore Funding Initiative and administered by the IDM Programme Office through a grant for the joint Carnegie Mellon/Singapore Management University Living Analytics Research Centre.

AMS 2000 subject classifications. Primary 62F12, 91D30; secondary 62F30.

Key words and phrases. Degree sequence, differential privacy, β -model, existence of MLE, measurement error.

<p>This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in <i>The Annals of Statistics</i>, 2016, Vol. 44, No. 1, 87–112. This reprint differs from the original in pagination and typographic detail.</p>

1. Introduction and motivation. Random graph models whose sufficient statistics are degree sequences, d , such as the p_1 model for directed graphs or its special case, the β -model for undirected graphs [Holland and Leinhardt (1981), Chatterjee, Diaconis and Sly (2011), Olhede and Wolfe (2012), Rinaldo, Petrović and Fienberg (2013)] are commonly used in modeling of real world networks. Although there is evidence that d alone does not capture all the structural information in a graph [e.g., Snijders (2003)], in many cases it is the only information available and every other structural property of a graph is estimated from random graph models based on d . In more general cases, random graph models based on d serve as a natural starting point for modeling networks; they may also serve as null models for hypothesis testing [Perry and Wolfe (2012), Zhang and Chen (2013)]. However, the degrees may carry confidential and sensitive information, and thus limit our ability to share such data more widely for the purpose of statistical inference. For example, in epidemiological studies of sexually transmitted disease [e.g., see Helleringer and Kohler (2007)], a survey collects information on the number of sexual partners of an individual, which provides an estimate of the degree of each node that is then used for modeling and reconstruction of a sexual network. The benefits of analyzing such networks are clear [e.g., Goodreau, Kitts and Morris (2009)], but releasing such sensitive information raises significant privacy concerns [e.g., (Narayanan and Shmatikov (2009))].

Data privacy is a growing problem due to the large amount of data being collected, stored, analyzed and shared across multiple domains. Statistical Disclosure Control (SDC) aims at designing data sharing mechanisms that address the trade-off between minimizing the risk of disclosing sensitive information and maximizing of data utility; for more details on SDC methodology, see, for example, Willenborg and de Waal (1996), Fienberg and Slavković (2010), Ramanayake and Zayatz (2010) and Hundepool et al. (2012). More recently, data privacy research has evolved with a focus on designing mechanisms that satisfy some *rigorous* notions of privacy but at the same time provide meaningful utility.

Differential Privacy (DP) [Dwork et al. (2006a)] has emerged as a key rigorous definition of privacy and as a way to inform the design of privacy mechanisms with pre-specified worst case disclosure risk. However, existing DP mechanisms are designed with a focus on estimating accurate summary statistics of the data, as opposed to estimating parameters of a model that are consistent and have correct confidence intervals; see Smith (2008) and Vu and Slavković (2009) for exceptions. As recently shown by Duchi, Jordan and Wainwright (2013), estimating parameters of models (that correspond to population quantities) and estimating summary statistics are fundamentally different problems, especially in the privacy context. However, the privacy mechanism is typically ignored and the perturbed statistics are used for subsequent analyses. Among many potential problems, ignoring the privacy

mechanism can lead to invalid even nonexistent parameter estimates, as initially demonstrated in Fienberg, Rinaldo and Yang (2010), Karwa and Slavković (2012) and in this paper.

This paper addresses the above mentioned fundamental problem of performing valid statistical inference using data released by a differentially private mechanism. Our work demonstrates that to obtain optimal parameter estimates by using data shared by privacy preserving mechanisms, new estimation procedures must be derived for specific classes of inference problems by modeling the privacy mechanism as a nonlinear measurement error process. The nonlinearity arises from the fact that noise is usually added to the sufficient statistics, as opposed to the data [see also Carroll et al. (2006)]. We illustrate the proposed principles in the context of special but important case of sharing network data using differential privacy; however, these principles are applicable beyond the specific privacy mechanism and the models considered here.

For network data, DP comes in two variants: Edge Differential Privacy [e.g., see Nissim, Raskhodnikova and Smith (2007)] and Node Differential Privacy [e.g., see Kasiviswanathan et al. (2013)], designed to limit disclosure of edge and node (along with its edges) information, respectively, in a graph G . We focus on the edge differential privacy with a goal of estimating the parameters of the β -model of random graphs whose sufficient statistics are network's degrees d . One of the popular ways of releasing d (and in general any summary statistic) to protect privacy is to release $z = d + e$, where e is some noise. In some cases, z is post-processed to reduce error [e.g., see Hay et al. (2009) for release of degree partitions] with the end goal to obtain an approximate estimate of the summary statistic of the data. However, the end goal of a statistical inference is not the estimation of statistics, in fact, the sufficient statistics are the starting point. Without any additional tools, the analyst is forced to directly use the noisy summary statistic z for inference. We present techniques to take into account the noise addition process and thereby consistently compute the maximum likelihood estimates (MLE) of the β -model from a noisy degree sequence. The following are the more specific contributions of this paper:

1. In Theorem 1, we derive necessary and sufficient conditions for the existence of MLE of the β -model, a result applicable beyond the privacy context. These conditions are computationally more efficient than those of Rinaldo, Petrović and Fienberg (2013), which are more general, but computationally intractable. This result gives insights into the conditions when the parameter estimates do not exist due to noisy statistics arising from privacy or possibly from sampling and censoring [Handcock and Gile (2010)].
2. Using the result on existence of MLE, we illustrate that ignoring the privacy mechanism and directly using the noisy statistic z for inference may

lead to issues such as nonexistence of MLE of the β -model. We also illustrate that the customary practice of simply minimizing the L_1 and/or L_2 distance between original and noisy statistics are not sufficient measures to guarantee statistical utility, and thus a valid inference. In particular, to obtain optimal and valid parameter estimates, the privacy mechanism must be explicitly taken into account when estimating the sufficient statistics from their noisy versions.

3. By modeling the privacy mechanism as a (known) measurement error process, we obtain a private maximum likelihood estimate \hat{d} of the degree sequence d , from its noisy counterpart z . In Theorem 2 and Algorithm 2, we show that this estimation problem can be solved efficiently, using a well-known characterization of degree sequences due to Havel (1955) and Hakimi (1962). This is a nonstandard maximum likelihood estimation problem where the parameter set is discrete and its dimensionality increases with the sample size. Using simulation studies, we show that \hat{d} has smaller error and greater statistical utility when compared to using z directly for parameter estimation.

4. In Theorems 3 and 4, we derive a differentially private consistent and asymptotically normal estimator $\hat{\beta}_\varepsilon$ of the parameters of the β -model of random graphs, by using the proposed estimated \hat{d} (instead of z). $\hat{\beta}_\varepsilon$ then can be used to generate valid synthetic graphs. Consistency of the usual MLE of β , without any privacy constraints, was shown by Chatterjee, Diaconis and Sly (2011) and its asymptotic normality was established in Yan and Xu (2013). Critically, since the proposed $\hat{\beta}_\varepsilon$ achieves the same rate as the nonprivate estimator, we show that asymptotically privacy comes at no additional cost in this setting.

The rest of the paper is organized as follows. In Section 2, we introduce the notation and the key results on the existence of MLE of the β -model and inference from noisy statistics. In Section 3, we describe our privacy model. Section 4 forms the core of the paper where we present our main results on estimating differentially private parameters of the β -model and on generating synthetic graphs. In Section 5, we extend our algorithm to release degree partitions and compare it to that of Hay et al. (2009). In Section 6, we evaluate our proposed estimators on real graphs. In Section 7, we briefly discuss avenues for future work, including the challenges in extending our work to larger class of β -models. Proofs are presented in Section 8 and the supplementary material [Karwa and Slavković (2015)].

2. Statistical inference with degree sequences. Let G_n denote a *simple, labeled* undirected graph on n nodes and let m be the number of edges in the graph. Let V be the vertex set and E be the edge set of the graph. A *simple* graph is a graph with no self-loops and multiple edges, that is,

for any $i \in [n]$, $(i, i) \notin E$, and $|\{(i, j) : (i, j) \in E\}| = 1$. A *labeled* graph is a graph with a fixed ordering on its nodes, that is, there is a fixed mapping from V to $\{1, \dots, n\}$. All the graphs considered in this paper are simple and undirected. Let \mathcal{G} denote the set of all such graphs. The distance between two graphs G and G' is defined as the number of edges on which the graphs differ and is denoted by $\delta(G, G')$. G and G' are said to be neighbors of each other if the distance between them is at most 1. The degree d_i of a node i is the number of nodes connected to it.

DEFINITION 1 (Degree sequence and degree partition). Consider a labeled graph with label $\{1, \dots, n\}$. The degree sequence of a graph d is defined as the sequence of degrees of each node, that is, $d = \{d_1, \dots, d_n\}$. The degree sequence ordered in nonincreasing order is called the degree partition and is denoted by \bar{d} , that is, $\bar{d} = \{d_{(1)}, \dots, d_{(n)}\}$ where $d_{(i)}$ is the i th largest degree.

Given a degree sequence d , there can be more than one graph with different edge-sets E , but the same degree sequence d . Each such graph is called a realization of d . Let $\mathcal{G}(d)$ be the set of simple graphs on n vertices with degree sequence d . Not every integer sequence of length n is a degree sequence. Sequences that can be realized by a simple graph are called *graphical degree sequences*. Graphical degree sequences have been studied in depth and admit many characterizations. One of the characterizations called the Havel–Hakimi criteria, due to Havel (1955) and Hakimi (1962), is central to the proof of Algorithm 2 that estimates a graphical degree sequence from the noisy sequence z ; see the proof of Theorem 2 in the supplementary material [Karwa and Slavković (2015)] for the statement of the characterization. We denote the set of all graphical degree sequences of size n by DS_n and the set of all graphical degree partitions of size n by DP_n .

2.1. Statistical inference with the β -model. One of the simplest random graph models involving the degree sequence is called the β -model, a term coined by Chatterjee, Diaconis and Sly (2011). We can describe this model in terms of independent Bernoulli random variables. Let $\beta = \{\beta_1, \dots, \beta_n\}$ be a fixed point in \mathbb{R}^n . For a random graph on n vertices, let each edge between nodes i and j occur independently of other edges with probability

$$p_{ij} = \frac{e^{\beta_i + \beta_j}}{1 + e^{\beta_i + \beta_j}},$$

where $\{\beta_1, \dots, \beta_n\}$ is the vector of parameters.

This model admits many different characterizations. For example, it arises as a special case of p_1 models [Holland and Leinhardt (1981)] and a log-linear model [Rinaldo, Petrović and Fienberg (2013)]. It is also a special case of

the discrete exponential family of distributions on the space of graphs when the degree sequence is a sufficient statistic. Thus, if G is a graph with degree sequence $\{d_1, \dots, d_n\}$, then the β -model is described by

$$P(G = g) \propto \exp \sum_{i=1}^n d_i \beta_i.$$

We can also consider a version of the β -model where the degree partition \bar{d} is a sufficient statistic. Such a model may be used if the ordering of the nodes is irrelevant.

In modeling real world networks, there are two very common inference tasks associated with the β -model:

1. Sample graphs from $\mathcal{U}(d)$ —the uniform distribution over the set of all graphs with degree sequence d .
2. Estimate parameters of the β -model using d and generate synthetic graphs from the β -model.

These tasks are useful, for example, in modeling network when the degree sequence is the only available information [Helleringer and Kohler (2007)], and in performing goodness-of-fit testing of more general network models [Hunter, Goodreau and Handcock (2008)]. A natural question to ask is under what conditions on d and \bar{d} are these two tasks possible: (a) Under what conditions does the MLE of the β -model exist? and (b) When is it possible to sample from $\mathcal{U}(d)$? In the next section, we study the conditions on d and \bar{d} that allow us to perform these inference tasks.

2.2. Existence of MLE of the β -model. Let $\hat{\beta}(d)$ denote the maximum likelihood estimate of β obtained using d . If we consider the degree partition version of the β -model, the MLE is denoted by $\hat{\beta}(\bar{d})$. From the properties of exponential families, it follows that $\hat{\beta}(d)$ must satisfy the following moment equations:

$$(2.1) \quad d_i = \sum_{j \neq i} \frac{e^{\hat{\beta}_i + \hat{\beta}_j}}{1 + e^{\hat{\beta}_i + \hat{\beta}_j}}.$$

A solution to these equations can be obtained in many ways. Most of them require iterative procedures [Hunter (2004), Chatterjee, Diaconis and Sly (2011)]. These procedures do not converge, or may converge to a meaningless value, when the MLE does not exist.

In Theorem 1, we describe necessary and sufficient conditions for existence of the MLE of the β -model. These conditions lead to an $O(n^2)$ algorithm to check for the existence of the MLE for the degree sequence β -model and show that the MLE never exists for the degree partition β -model. To the

best of our knowledge, this is the first efficient algorithm for checking the existence of MLE of the β -model. The proof of Theorem 1 is in Section 8.1.

From the theory of exponential families [Barndorff-Nielsen (1978)], it follows that $\hat{\beta}(d)$ exists if and only if d lies in the relative interior of convex hull of DS_n . Although the facets of $\text{Conv}(DS_n)$ are completely characterized in Mahadev and Peled (1995), one cannot use the linear inequality description of $\text{Conv}(DS_n)$ to check if d lies in the relative interior. This is because $\text{Conv}(DS_n)$ is a complex combinatorial object and the number of facet defining inequalities [given in equation (8.1)] are at least exponential in n . Rinaldo, Petrović and Fienberg (2013) use results from the existence of MLE of discrete exponential families [Rinaldo, Fienberg and Zhou (2009)] to devise an algorithm to check for the existence of MLE in what they refer to as a *generalized* β -model. Their algorithm is based on the so-called ‘‘Cayley embedding’’ which is a reparametrization of the β -model as a log-linear model. Although general, their algorithm works only for graphs up to a few hundreds of nodes, and its computational complexity is unknown.

The key technique that we use for proving Theorem 1 is to study an ‘‘asymmetric’’ part of $\text{Conv}(DS_n)$. Specifically, we work with $\text{Conv}(DP_n)$, the convex hull of degree partitions, instead of $\text{Conv}(DS_n)$. Intuitively, $\text{Conv}(DP_n)$ can be considered as a ‘‘asymmetrized’’ version of $\text{Conv}(DS_n)$ —every permutation equivalent degree sequence is mapped to a single degree partition [see also Bhattacharya, Sivasubramanian and Srinivasan (2006)]. This asymmetrization, remarkably, allows us to characterize the boundary of $\text{Conv}(DS_n)$, and at the same time, greatly reduce the computational complexity. We conjecture that this technique of asymmetrizing a polytope can be extended to other discrete exponential families to derive efficient algorithms that characterize their boundary.

THEOREM 1. *Let G be a graph. Let d be its degree sequence and \bar{d} be the corresponding degree partition obtained by ordering the terms of d in a nonincreasing order. Consider the following set of inequalities:*

$$(2.2) \quad \begin{aligned} &\bar{d}_i > 0 \quad \text{and} \quad \bar{d}_i < n - 1 \quad \forall i \quad \text{and} \\ &\sum_{i=1}^k \bar{d}_i - \sum_{i=n-l+1}^n \bar{d}_i < k(n-1-l) \quad \text{for } 1 \leq k+l \leq n, \end{aligned}$$

$$(2.3) \quad \bar{d}_{i+1} - \bar{d}_i < 0 \quad \text{for } i = 1 \text{ to } n.$$

The following statements are true:

1. *The MLE of the degree partition β -model $\hat{\beta}(\bar{d})$ exists iff \bar{d} satisfies the system of inequalities in (2.2) and (2.3). In particular, the MLE for the degree partition β -model never exists.*

2. If the MLE of the degree sequence β -model $\hat{\beta}(d)$ exists, then \bar{d} satisfies the system (2.2).

3. If \bar{d} satisfies the system (2.2), then $\hat{\beta}(d)$ exists for any $d = \pi\bar{d}$ where π is a permutation on $\{1, \dots, n\}$.

REMARKS.

1. The system of inequalities in equation (2.2) are central to the results of Theorem 1. There are only $O(n^2)$ inequalities to check, as opposed to exponentially many inequalities that describe $\text{Conv}(DS_n)$. Thus, an important practical consequence of this result is the first quadratic time algorithm to detect the boundary points of $\text{Conv}(DS_n)$ and check for the existence of MLE of the degree sequence β -model.

2. Statement 3, the converse condition in Theorem 1 is stronger than statement 2. It implies that if \bar{d} satisfies the system (2.2), then the MLE of β computed using any permutation of \bar{d} exists.

3. Theorem 1 does not imply that d is in $\text{ri}(\text{Conv}(DS_n))$ if and only if \bar{d} is in $\text{ri}(\text{Conv}(DP_n))$. In fact, this is not true—no (graphical) degree partitions exists in the relative interior of $\text{Conv}(DP_n)$; all degree partitions lie on at least one of the boundaries defined by equation (2.3).

4. When we observe a single graph, the MLE for the degree partition β -model never exists. From this point onward, we will use the term “MLE of β ” to mean the MLE of the degree sequence β -model, even when using a degree partition, since every degree partition is also a degree sequence.

5. The degree distribution is the histogram of degree partition, and furthermore the degree distribution and the degree partition are one to one transformations of each other, one can be obtained from the other via a nonlinear transformation. Most recently, Sadeghi and Rinaldo (2014) show that the MLE of the degree distribution model also never exists which complements our results on the degree partition.

2.3. *Sampling from $\mathcal{U}(d)$.* Sampling graphs from the set $\mathcal{U}(d)$ is possible only if the set $\mathcal{G}(d)$ is nondegenerate. Moreover, for there to exist a nontrivial probability distribution on this set, its cardinality should be greater than 1. Proposition 1 presents sufficient conditions on d under which this is true; the proof appears in Section II of the supplementary material [Karwa and Slavković (2015)].

PROPOSITION 1. *Let d be a sequence of real numbers. Consider the set $\mathcal{G}(d)$, the set of all simple graphs with degree sequence equal to d . If d is a point in DS_n , and if d lies in the relative interior of $\text{Conv}(DS_n)$, then $|\mathcal{G}(d)| > 1$.*

2.4. *Inference using noisy statistics.* Theorem 1 and Proposition 1 give sufficient conditions for estimating parameters of the β -model and for sampling from the space of related graphs. However, in many real world applications, the exact degree sequence d of a graph is not available. Instead, we observe a “noisy” sequence z either due to sampling issues or due to privacy constraints. Corollary 1 gives sufficient conditions for obtaining valid inference in the β -model when using such “noisy” sequences.

COROLLARY 1. *Let z be any sequence of integers of length n . Consider the following two inference tasks: (1) Estimating the MLE of β -model using z . (2) Sampling from the set $\mathcal{U}(z)$. A sufficient condition to ensure that the MLE exists and $\mathcal{U}(z)$ is nonempty is that z is a point in DS_n and lies in the relative interior of convex hull of DS_n .*

In Section 4, we consider the case where z is a noisy degree sequence obtained by applying a differentially private mechanism to d . We discuss in more detail why directly using z instead of d typically leads to invalid inference and apply the results of this section to obtain valid statistical inference by finding an estimate of d that satisfies conditions of Corollary 1.

3. Edge differential privacy. Differential privacy has become one of the most popular models of reasoning formally about privacy. In a typical interactive setting, data users can ask *queries* about the data, which can be in the form of sufficient statistics, and they would receive back differentially private answers. This type of a privacy mechanism can be formalized as a family of conditional probability distributions, which define a distribution on the answers, conditional on the data; for a statistical overview of differential privacy; see Wasserman and Zhou (2010).

In this paper, we focus on edge differential privacy (EDP) where the goal is to protect the topological information of the graph. EDP is defined to limit disclosure related to presence or absence of edges in a graph (or relationships between nodes) as the following definition illustrates.

DEFINITION 2 (Edge differential privacy). Let $\varepsilon > 0$. A randomized mechanism (or a family of conditional probability distributions) $\mathcal{Q}(\cdot|G)$ is ε -edge differentially private if

$$\sup_{G, G' \in \mathcal{G}, \delta(G, G')=1} \sup_{S \in \mathcal{S}} \log \frac{\mathcal{Q}(S|G)}{\mathcal{Q}(S|G')} \leq \varepsilon,$$

where \mathcal{S} is the set of all possible outputs (or the range of \mathcal{Q}).

ε is the privacy parameter that, as we see below, controls the amount of noise added to a statistic; small value of ε means more privacy protection, but leads to larger noise in the statistic being released. Roughly, EDP

requires that any output of the mechanism \mathcal{Q} on two neighboring graphs should be close to each other. Along the lines of Theorem 2.4 in Wasserman and Zhou (2010), one can show that EDP makes it nearly impossible to test the presence or absence of an edge in the graph, thus providing protection.

The most common mechanism to release the output of any statistic f under differential privacy is the Laplace mechanism [e.g., see Dwork et al. (2006a)] which adds continuous Laplace noise proportional to the *global sensitivity* of f .

DEFINITION 3 (Global sensitivity). Let $f : \mathcal{G} \rightarrow \mathbb{Z}^k$. The global sensitivity of f is defined as

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

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

Here, we propose to use a variant of this mechanism to achieve EDP by adding discrete Laplace noise, as described in Lemma 1, to the degree sequence of a graph (see Algorithm 1 in Section 4.1). Ghosh, Roughgarden and Sundararajan (2009) analyzed the discrete Laplace mechanism for one-dimensional counting queries and showed that it is universally optimal for a large class of utility metrics. The proof of Lemma 1 is given in Section I of the supplementary material [Karwa and Slavković (2015)].

LEMMA 1 (Discrete Laplace mechanism). Let $f : \mathcal{G} \rightarrow \mathbb{Z}^k$. Let Z_1, \dots, Z_k be independent and identically distributed discrete Laplace random variables with p.m.f. defined as follows:

$$P(Z = z) = \frac{1 - \alpha}{1 + \alpha} \alpha^{|z|}, \quad z \in \mathbb{Z}, \alpha \in (0, 1).$$

Then the algorithm which on input G outputs $f(G) + (Z_1, \dots, Z_k)$ is ε -edge differentially private, where $\varepsilon = -GS(f) \log \alpha$.

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

LEMMA 2 [Dwork et al. (2006b), 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 using Lemma 2, we can ensure that any post-processing done on the noisy degree sequences obtained as an output of a differentially private mechanism is also differentially private. In particular, this means that applying the proposed Algorithm 2 to the output of a differentially private mechanism also preserves differential privacy.

Algorithm 1 *Input:* A graph G and privacy parameter ε . *Output:* Differentially private answer to the degree sequence of G

```

1: Let  $d = \{d_1, \dots, d_n\}$  be the degree sequence of  $G$ 
2: for  $i = 1 \rightarrow n$  do
3:   Simulate  $e_i$  from discrete Laplace with  $\alpha = \exp(-\varepsilon/2)$ 
4:   Let  $z_i = d_i + e_i$ 
5: end for
6: return  $z = \{z_1, \dots, z_n\}$ 

```

4. Estimating parameters of the β -model using noisy degree sequences and releasing synthetic graphs. In this section, we present our main results on obtaining consistent and asymptotically normal differentially-private MLEs for the β -model. These results support two main objectives: (1) To achieve statistical inference that is both optimal and private for the β -model, and (2) to release synthetic graphs from the β -model in a differentially private manner.

Our approach is based on three steps. In the first step, we release the degree sequence, which is a sufficient statistic of the β -model, using the discrete Laplace mechanism described in Lemma 1. In the second step, we model the Laplace mechanism as a measurement error on the sufficient statistics and “de-noise” the noisy sufficient statistic by using maximum likelihood estimation. In the third step, the de-noised sufficient statistic is used to estimate the parameters of the β -model from which synthetic graphs can be generated. Since each of these steps uses only the output of a differentially private algorithm, by Lemma 2, the generated synthetic graphs are also differentially private. Step 2 of modeling the privacy mechanism as a measurement error process and re-estimating the degree sequence is critical, as we show in the proofs of Theorems 3 and 4, since it allows the third step to produce consistent and asymptotically normal parameter estimates. In the next subsections, we look at each of these steps in detail and describe the associated algorithms and theoretical results.

4.1. *Releasing the degree sequence privately.* Since the degree sequence d (or degree partition \bar{d}) is a sufficient statistic of the β -model, the first step releases these statistics under differential privacy via Algorithm 1. We use the discrete Laplace mechanism (Lemma 1). The global sensitivity of both d and \bar{d} is 2 since adding or removing an edge can change the degree of at most two nodes, by 1 each.

Can we use z , a differentially private output of the degree sequence d released by Algorithm 1, directly for inference and generate synthetic graphs? Most work on differential privacy advocates using z or some post-processed form of z as a “proxy” of d for inference. This, however, ignores the noise

addition process. Furthermore, a more serious issue is that z may not satisfy the conditions of Corollary 1.

To understand how z fails the conditions of Corollary 1, consider task (1) from Section 2 where the goal is to simulate random graphs from the $\mathcal{U}(d)$ by using the output z instead of d . Recall that $\mathcal{U}(d)$ is nonempty if and only if d is a point in DS_n , that is, d is a graphical sequence. What are the chances that z is graphical? If z is a sequence of positive integers, the chances are asymptotically at best 50%; see Arratia and Liggett (2005). In the present case, z is supported on the set of integers, \mathbb{Z}_n as it is obtained by adding discrete Laplace noise to d . Hence, it is quite unlikely for z to even be in $\text{Conv}(DS_n)$. Thus, in many cases z cannot be used directly to perform task (1).

How about task (2) of estimating β ? Let $\hat{\beta}(d)$ denote the MLE of β obtained using d . A basic requirement is the following: If $\hat{\beta}(d)$ exists, then $\hat{\beta}(z)$, should also exist. As we mentioned, the existence of MLE is guaranteed only if z lies in the interior of convex hull of DS_n . As discussed earlier, even if d lies in the interior of convex hull of DS_n , z need not. Thus, directly inputting z into a procedure that estimates the MLE may lead to meaningless results as the MLE may not exist. See also, Figure 1 in Section 5 for an empirical demonstration of nonexistence of MLE when using z to estimate the parameters.

In the next section, we will see that these issues can be resolved by modeling the privacy mechanism as a measurement error process, and computing an estimate \hat{d} of d , from the noisy sequence z , that satisfies the conditions in Corollary 1 with very high probability. Thus, one of the advantages of using \hat{d} (instead of z) for estimation ensures that the MLE of β exists; see Theorem 3 for a precise statement. In fact, when using \hat{d} for estimation, not only does the MLE exist, but the MLE is consistent and asymptotically normally distributed, as proved in Section 4.3.

4.2. Maximum likelihood estimation of degree sequence. We model the privacy mechanism from Algorithm 1 as a measurement error on the degree sequence, and use maximum likelihood estimation to “de-noise” the noisy sequence z . The noise addition process here is regarded as special type of measurement error since we know the exact distribution of the error. Hence, despite of the fact that we observe a single sample from the measurement error process (the degree sequence is released only once), we can recover an estimate of the original sequence. This takes the privacy mechanism into account in a principled manner and leads to an estimate of d that can then be used for inference. More formally, the output of Algorithm 1 generates n random variables z_i , such that $z_i = d_i + e_i$ where $e_i \sim \text{DLap}(\alpha)$, for $i = 1$ to n and $d = \{d_1, \dots, d_n\} \in DS_n$. Note that α is known and we treat d as the

Algorithm 2 *Input:* A sequence of integers z of length n . *Output:* A graph G on n vertices with degree sequence \hat{d}

```

1: Let  $G$  be the empty graph on  $n$  vertices
2: Let  $S = \{1, \dots, n\}$ 
3: while  $|S| > 0$  do
4:    $S = S \setminus T$  where  $T = \{i : z_i \leq 0\}$ 
5:   Let  $\text{pos} = |S|$ 
6:   Let  $z_{i^*} = \max_{i \in S} z_i$ . Let  $i^* = \min\{i \in S : z_i = z_{i^*}\}$  and let  $h_{i^*} = \min(z_{i^*}, \text{pos} - 1)$ 
7:   Let  $\mathcal{I} =$  indices of  $h_{i^*}$  highest values in  $z(S \setminus \{i^*\})$  where  $z(S)$  is the sequence  $z$  restricted to the index set  $S$ 
8:   Add edge  $(i^*, k)$  to  $G$  for all  $k \in \mathcal{I}$ 
9:   Let  $z_i = z_i - 1$  for all  $i \in \mathcal{I}$  and  $S = S \setminus \{i^*\}$ 
10: end while
11: return  $G$ 

```

fixed unknown parameter in DS_n . We propose Algorithm 2 that produces the maximum likelihood estimator \hat{d} of d from the vector of noisy degrees z , and Theorem 2 asserts its correctness. The proof of Theorem 2 is deferred until Section IV of the supplementary material [Karwa and Slavković (2015)].

THEOREM 2 (MLE of degree sequence). *Let $z = \{z_i\}$ be a sequence of integers of length n obtained from Algorithm 1. The degree sequence of graph G produced by Algorithm 2 is a maximum likelihood estimator of d .*

Here, we make some remarks on the complexity of this key result. Note that the measurement error model and the corresponding maximum likelihood estimation of the degree sequence is nonstandard—the number of parameters to be estimated $(d_i, i = 1, \dots, n)$ is equal to the number of observations $(z_i, i = 1, \dots, n)$, and the parameter space is discrete and very large—the convex hull of the parameter set is full dimensional for $n \geq 4$. Computing an MLE of d in the measurement error model is equivalent to finding a L_1 “projection” of z on DS_n , that is, finding a graphical degree sequence in DS_n closest to z in terms of the L_1 distance:

$$(4.1) \quad \hat{d} = \underset{h \in DS_n}{\operatorname{argmin}} \|h - z\|_1.$$

Here, the parameter set DS_n is a collection of points, and it admits several characterizations. We found the Havel–Hakimi characterization to be the most useful in producing an efficient procedure for estimating the MLE, as evident in the proof of Theorem 2; see Section IV of Karwa and Slavković (2015). In fact, a careful analysis of Algorithm 2 shows that it is a modified Havel–Hakimi procedure applied to the noisy sequence z .

The Havel–Hakimi algorithm is a “certifying” algorithm in that it produces a certificate that a degree sequence is graphical, that is, if the input to the algorithm is a (graphical) degree sequence, it outputs a graph that realizes it. Remarkably, our proof of Theorem 2 shows that we can convert such a certifying algorithm into an algorithm (e.g., Algorithm 2) that performs L_1 “projection” on the set DS_n . We conjecture that our proof techniques apply to more general polytopes such as the polytope of degree sequences of bipartite graphs or directed graphs. In cases where a certifying algorithm like the Havel–Hakimi is available for these polytopes, our proof techniques can be used to devise algorithms for L_1 optimization over the corresponding set of graphical degree sequences.

Even though the maximum likelihood estimation is equivalent to an L_1 projection, there are many differences from the traditional projection. The set DS_n has “holes” in it and is not a convex set. As an example, every point whose L_1 norm is not divisible by 2 is not included in the set. Due to this, the L_1 projection need not be on the boundary of the convex hull of DS_n . Moreover, there can be more than one degree sequence that attains the optimal L_1 distance. Thus, the MLE of d is actually a set and Algorithm 2 finds a point in this set. Specifically, the following is true.

LEMMA 3. *Let d^* be the output of Algorithm 2. Let $Z = \{i : d_i^* = 0 \text{ and } z_i < 0\}$ and $P = \{i : d_i < z_i \text{ and } d_i > 0\}$, and let $|P| \neq 0$. Let $k \in Z$. Then there exists a degree sequence d such that $d_k > 0$ and $\|d^* - z\|_1 = \|d - z\|_1$.*

Lemma 3 [proof of which is in Section III of Karwa and Slavković (2015)] shows that the de-noised degree sequence is not unique. Hence, the noise addition process provides privacy as the original degree cannot be recovered exactly. Another way to interpret this result is that the Laplace noise adds more noise than what is needed to ensure differential privacy, and Algorithm 2 “removes” this additional noise, since applying Algorithm 2 does not degrade privacy, but crucially improves utility.

Note that Algorithm 2 is efficient and it runs in time $O(n \log n + m)$ where n is the number of nodes and m is the number of edges. Algorithm 2 returns a graph G whose degree sequence is \hat{d} , thus, by definition, \hat{d} is graphical. By randomizing G , for example, by using the techniques in Blitzstein and Diaconis (2010) or Ogawa, Hara and Takemura (2011), the output from Algorithm 2 can also be used to generate synthetic graphs from the uniform distribution of graphs with a fixed degree sequence, $\mathcal{U}(d)$.

In some cases, especially when some of the z_i ’s are negative, G may be a disconnected graph. In such cases, whenever the conditions of Lemma 3 are satisfied, we use it to modify the optimal degree sequence so that it corresponds to a connected graph. (Note that being the degree sequence of a connected graph does not ensure that the MLE exists, but the opposite is

true—the MLE of β does not exist if the degree sequence is realized by a disconnected graph.) The proof of Lemma 3 in Section III of the supplementary material gives the steps for the construction of the modified sequence. It is easy to see that verification of the conditions of Lemma 3 and the construction of the modified sequence takes $O(n \log n)$ time. Hence, asymptotically, this step does not increase the computational complexity of Algorithm 2. We now proceed to the task of estimating β using \hat{d} .

4.3. *Asymptotic properties of the private estimate of β .* Let \hat{d} denote the ε -differentially private estimate of d obtained by using Algorithms 1 and 2. A private MLE of β can be obtained by plugging \hat{d} in the maximum likelihood equations (2.1) and solving for β ; let us denote this estimate by $\hat{\beta}(\hat{d})$. Since \hat{d} is ε -differentially private, by Lemma 2, $\hat{\beta}(\hat{d})$ is also ε -differentially private. But how does $\hat{\beta}(\hat{d})$ compares to the estimate $\hat{\beta}(d)$ obtained from the original degree sequence d ? We demonstrate the utility of the proposed private estimate of β by proving two key results in Theorems 3 and 4, that is, $\hat{\beta}(\hat{d})$ is consistent and asymptotically normal.

Consistency—Consistency of the maximum likelihood estimator of β in the nonprivate case was shown by Chatterjee, Diaconis and Sly (2011). Here, we show that our proposed private estimator of β is also consistent, that is one can consistently estimate the parameters of the β -model using \hat{d} (as opposed to using d).

Theorem 3 shows that using \hat{d} to estimate the MLE guarantees both the existence of MLE and the uniform consistency (in contrast to naively using the differentially private output z that does not even guarantee that the MLE exists as discussed in Sections 4.1 and 4.2).

THEOREM 3 (Asymptotic consistency). *Let G be a random graph from the β -model and let $d = (d_1, \dots, d_n)$ be its degree sequence. Let $L = \max_i |\beta_i|$. Let $\hat{d} = (\hat{d}_1, \dots, \hat{d}_n)$ be the differentially private maximum likelihood estimate of d obtained from output of the Algorithm 2, and let*

$$\hat{d}_i = \sum_{j \neq i} \frac{e^{\hat{\beta}_i + \hat{\beta}_j}}{1 + e^{\hat{\beta}_i + \hat{\beta}_j}}$$

be the maximum likelihood equations. Let $C(L)$ be a constant that depends only on L . Then for $\varepsilon_n = \Omega(\frac{1}{\sqrt{\log n}})$, there exists a unique solution $\hat{\beta}(\hat{d})$ to the maximum likelihood equation such that

$$\mathbb{P}\left(\max_i |\hat{\beta}_i(\hat{d}) - \beta_i| \leq C(L) \sqrt{\frac{\log n}{n}}\right) \geq 1 - C(L)n^{-2}.$$

The proof of Theorem 3 is given in Section V of the supplementary material [Karwa and Slavković (2015)]. This key result implies that asymptotically there is no cost to privacy in this setting in relation to obtaining valid inference. In particular, the result shows that for large n and $\varepsilon = \Omega(\frac{1}{\sqrt{\log n}})$, the MLE of β obtained from \hat{d} exists and is unique and can be estimated with uniform accuracy in all coordinates. In practice, the dependence of ε on n can be improved by numerically computing and checking if the tail bound in Lemma C in the supplementary material [Karwa and Slavković (2015)], needed for the proof of Theorem 3, is satisfied. Thus, this theorem gives practical guidelines on whether for a given ε and n combination, the consistency result holds.

Finally, we want to point that if one is allowed to release d many times using Algorithm 1, one can average out the noise due to the Laplace mechanism and get consistency trivially by using the law of large numbers. This is not allowed, as the privacy loss of each release is additive in terms of ε and would defeat the purpose of privacy. Hence, to provide meaningful privacy, the sample size of the private degree sequence is 1, that is, d is released only once using the Laplace mechanism. Theorem 3 shows that consistency can still be obtained using a single private sample of the degree sequence.

Asymptotic normality—A central limit theorem for $\hat{\beta}(d)$ was derived in Yan and Xu (2013); see also Yan, Zhao and Qin (2015). In Theorem 4, we derive a similar central limit result for $\hat{\beta}(\hat{d})$. This distribution can be used to derive differentially private approximate confidence intervals and perform hypothesis tests on the parameter estimates. The proof is given in Section VI of the supplementary material [Karwa and Slavković (2015)].

Let the covariance matrix of $d = \{d_1, \dots, d_n\}$ be $V_n = \{v_{ij}\}$ where

$$v_{ij} = \frac{\exp \beta_i + \beta_j}{(1 + \exp \beta_i + \beta_j)^2}$$

and

$$v_{ii} = \sum_{j \neq i, j=1}^n v_{ij}.$$

THEOREM 4 (Asymptotic normality). *Let $L = \max_i |\beta_i|$ be a fixed constant and $\varepsilon = \Omega(\frac{1}{\sqrt{\log n}})$. Let \hat{d} be a differentially private maximum likelihood estimate of d obtained from Algorithm 2. Let $\hat{\beta}(\hat{d})$ be the MLE of the β -model obtained using \hat{d} . For any fixed $r \geq 1$, the random vector*

$$(\sqrt{v_{11}}(\hat{\beta}(\hat{d})_1 - \beta_1), \dots, \sqrt{v_{rr}}(\hat{\beta}(\hat{d})_r - \beta_r))$$

converges to a standard multivariate normal distribution.

5. Releasing graphical degree partitions. In this section, we extend Algorithm 2 to release degree partitions and compare it with previous work due to Hay et al. (2009).

One can release the degree partition \bar{d} instead of the degree sequence d in cases where the ordering of the nodes is not important, or one is interested in the degree distribution (histogram of degrees). The latter was the motivation of Hay et al. (2009) who instead of releasing the degree distribution, release the degree partition \bar{d} which has the same global sensitivity as d ; thus, Algorithm 1 can be used to release a noisy degree partition. Let z be the noisy answer, that is, $z = \bar{d} + e$. Hay et al. (2009) project z onto the set of integer partitions (nonincreasing integer sequences), which is a special case of isotonic regression (henceforth referred to as “Isotone”). They show that this reduces the L_2 error. Note, however, that the output need not be a graphical degree partition, that is, there may not exist any simple graph corresponding to the output.

To solve this issue, we propose using the following two step algorithm (referred to as “Isotone–Havel–Hakimi” or “Isotone–HH”) to release a graphical degree partition.

1. Let \bar{z} be the closest integer partition to z in terms of L_1 distance.
2. Let \hat{d} be the output of Algorithm 2 on input \bar{z} .

Unlike the case of degree sequence, this procedure does not estimate an MLE of \bar{d} . However, Corollary 2 shows that the estimate is still optimal in sense of the L_1 error, and more importantly, it is a point in DP_n that is closest to \bar{z} . The proof of Corollary 2 appears in Section VII of the supplementary material [Karwa and Slavković (2015)].

COROLLARY 2. *Let $\bar{z} = \{\bar{z}_i\}$ be a sequence of nonincreasing integers of length n . The degree partition of graph G output by Algorithm 2 on input \bar{z} is a solution to the optimization problem $\operatorname{argmin}_{h \in DP_n} \|h - \bar{z}\|_1$.*

Release of synthetic graphs here follows as discussed in Section 4.2.

6. Simulation results. In this section, we evaluate the finite sample properties of the differentially private estimator of β . We perform two sets of experiments. In the first set, we compare the utility of Hay et al. (2009) with our algorithm when releasing degree partitions \bar{d} . In the second set of experiments, we estimate β using the private estimate \hat{d} of Algorithm 2 and compare it with the estimates obtained by using the nonprivate degree d . We use three networks, two real and one simulated, described below.

1. Sampson Monastery Data [Sampson (1968)]—This is a real network of relationship between monks in a monastery. It consists of social relations

among a set of 18 monks. The original dataset was asymmetric and collected for three time periods. In this study, we symmetrize the network by using the upper triangular adjacency matrix of time period 1. There are 18 nodes and 35 edges in this network.

2. Karate Dataset [Zachary (1977)]—This is a real network of friendships between 34 members of a karate club at a US university in the 1970. It has 78 edges and 34 nodes.

3. Likoma Island [Helleringer and Kohler (2007)]—This is a simulated network of number of sexual partners of people living in the Likoma island. Helleringer et al. (2009) describes the study and data collection procedures based on a survey. Using the estimated degree sequence [obtained from the survey data and given in Helleringer et al. (2009)], we simulated a random network with the fixed degree sequence. The simulated network consists of 250 nodes and 248 edges.

Releasing \bar{d} to estimate β : The goal of these experiments is to compare *isotone* and *isotone-hh* algorithms for releasing differentially private versions of degree partitions \bar{d} . We evaluate these algorithms on two metrics. The first metric is the probability of the event R where $R = \{\hat{\beta}(y) \text{ exists}\}$, where y is output of the mechanism. The second metric is the median L_1 error between \bar{d} and y for fixed \bar{d} , that is, $\text{err}(\bar{d}) = \text{median}[|\bar{d} - y|]$. For each network and a fixed value of privacy parameter ε , \bar{d} is released $B = 500$ times using *isotone* and our *isotone-hh* procedure. Note that even though each release of \bar{d} is ε -edge differentially private, the entire simulation study is 500ε -edge differentially private. In practice, \bar{d} will be released only once. However, in the experiments, we are interested in evaluating the frequentist properties of the procedure, and hence we release the degree partition multiple times. Using these released degree partitions, we compute $P(R)$ and $\text{err}(\bar{d})$. This procedure is repeated for different levels of ε varying from 0 to 4, for all three datasets. Note that a larger ε means lower noise and less privacy. Figure 1 shows a plot of $P(E)$ and $\text{err}(\bar{d})$ normalized by the number of nodes for varying levels of ε .

As expected, for both algorithms, as ε increases, $P(R)$ increases and the median L_1 error decreases. In many cases, the MLE of the output of *isotone* fails to exist as it lies outside the convex hull of DP_n . $P(R)$ is significantly higher for *isotone-hh* for all three datasets. For instance for the Karate dataset, $P(R)$ quickly approaches 1 as ε increases, when using the *isotone-hh* algorithm, where as it never reaches 1 when using the *isotone* algorithm. The other two datasets exhibit similar behavior. We can also see that for the Likoma dataset, the gap between the two algorithms in terms of $P(R)$ is much higher when compared to the other two datasets. More specifically, when using the *isotone* algorithm, $P(R)$ increases slowly with ε for the Likoma dataset when compared to the other two datasets. On the

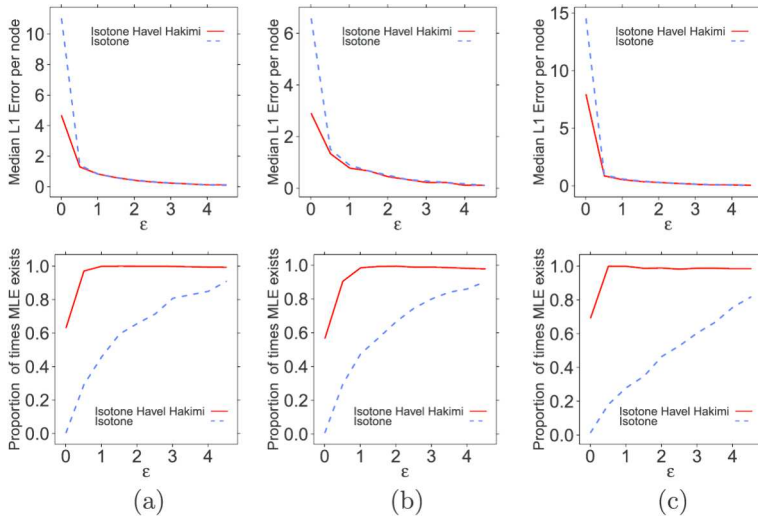


FIG. 1. Comparison of “Isotone” and “Isotone-HH” to release \bar{d} . The plots show the L_1 error and the probability that the MLE exists for varying levels of ε for three different networks. (a) Karate; (b) Sampson; (c) Likoma.

other hand, when using the *isotone-hh* algorithm, $P(R)$ increases quickly with ε for all three datasets. A possible explanation for the behavior of the *isotone* algorithm is that the Likoma data are sparse. Recall that $P(R)$ is 0 if the noisy sequence lies outside $\text{Conv}(DP_n)$ (see Theorem 1). Due to the sparsity of Likoma data, the degree partition is close to the boundary of $\text{Conv}(DP_n)$. In this case, adding Laplace noise puts the degree partition outside $\text{Conv}(DP_n)$, and the post-processing step of *isotone* is not sufficient to get a sequence inside $\text{Conv}(DP_n)$, and hence $P(R) = 0$ for such instances.

When considering the median L_1 error, the *isotone-hh* algorithm not only provides an increased probability that the MLE exists, but also provides more accurate estimates of \bar{d} , especially for smaller levels of ε . For instance, for $\varepsilon = 0.1$, for the Karate dataset, the median L_1 error per node in estimating the degree is 4 for the *isotone-hh* whereas it is greater than 10 for the *isotone* algorithm. Thus, we can see that *isotone-hh* offers more “utility” in terms of both estimating the MLE, and also in terms of the L_1 error.

Estimation of β using d : In the second set of experiments, we evaluate how close $\hat{\beta}(\hat{d})$ is to $\hat{\beta}(d)$. Here, $\hat{\beta}(d)$ is the estimate of β obtained by using the original degree sequence and $\hat{\beta}(\hat{d})$ is the estimate of β obtained by using the private degree sequence \hat{d} obtained from the output of Algorithm 2. Figure 2 shows a plot of the estimates of β on the y axis and degree on the x axis. The red line indicates $\hat{\beta}(d)$ and the green line indicates the median estimate of $\hat{\beta}(\hat{d})$. Also plotted are the upper (95th) and the lower (2.5th) quantiles of the

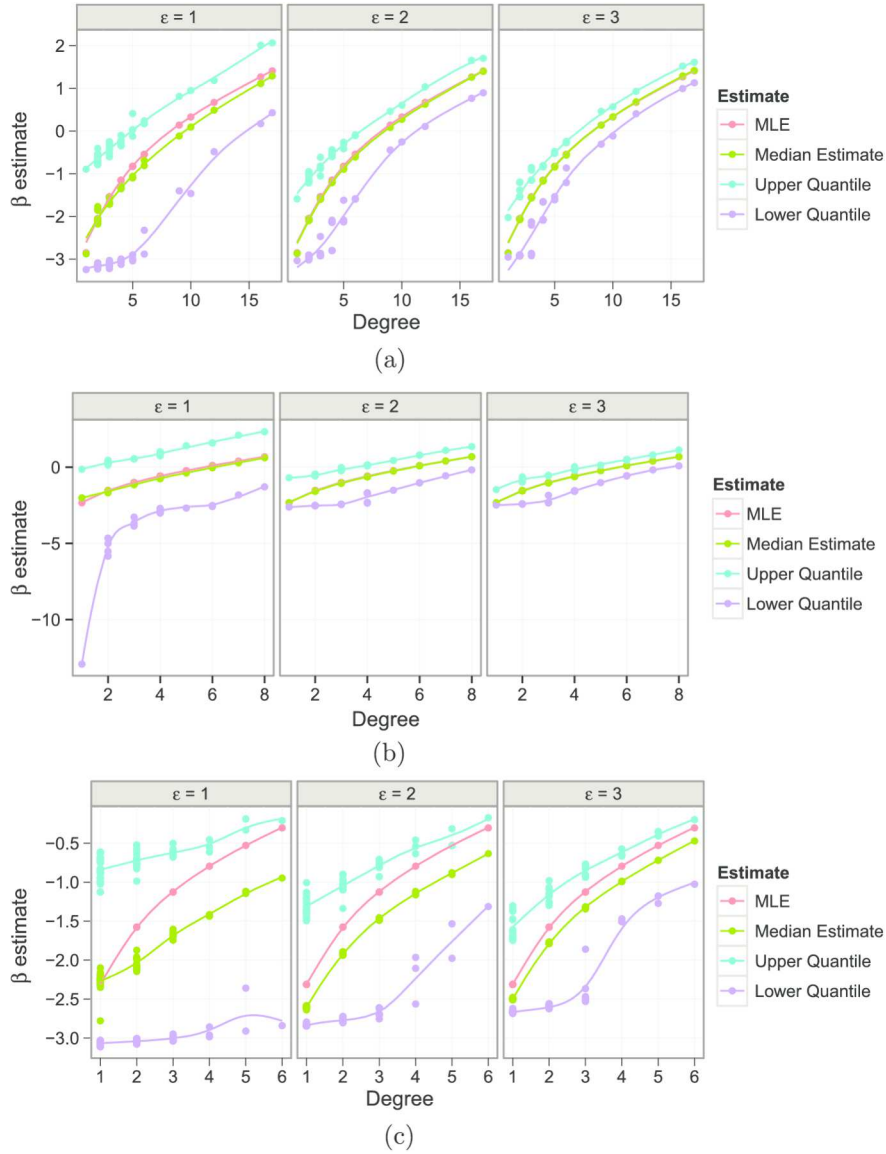


FIG. 2. Comparison of differentially private estimate of β with the MLE for three different datasets. The plots show the median and the upper (95th) and the lower (2.5th) quantiles. (a) Karate data; (b) Sampson data; (c) Likoma island data.

estimates. The results show that the median estimate of $\beta(\hat{d})$ is very close to $\beta(d)$ and lies within the 95 percent quantiles of the estimates. Moreover, as expected, as ϵ increases, the variance in the estimates get smaller. The median private estimates of β for the Karate and the Sampson dataset are

very close to the nonprivate MLE. However, the private estimates of β for the Likoma dataset have higher variance and are farther from MLE of $\beta(d)$ due to the fact that the Likoma graph is sparse and the β -model does not fit the original data very well. This suggests that the β -model may not be a *robust* model for sparse networks in the following sense. If the network is very sparse, the degree sequence of the original data may lie close to the boundary of $\text{Conv}(DS_n)$. Due to this, adding or removing a small number of edges may cause the degree sequence to end up being on the boundary.

7. Conclusions and future work. In this paper, we characterize the conditions for the existence of MLE of the degree partition and the degree sequence β -model that lead to an efficient quadratic time algorithm. Motivated by the privacy problem of sharing confidential data under rigorous privacy guarantees, that often falls short of satisfying data utility, we present techniques to perform valid and differentially private statistical inference with the β -model of random graphs and to release differentially private synthetic graphs from the β -model. We present an efficient maximum likelihood algorithm to re-estimate the original degree sequence from a noisy sequence released by a differentially private mechanism. We showed that this estimated degree sequence can be used to obtain a consistent and asymptotically normally distributed estimates of the parameters of the β -model, and thus incur no cost due to privacy from utility perspective. Using the example of the β -model, we showed that the noisy sufficient statistics z must be post-processed (or projected) in an appropriate manner by taking the noise mechanism into account in order to obtain optimal inference. In particular, by treating the privacy mechanism as a nonlinear measurement error model, one can estimate the sufficient statistics from their noisy counterparts and obtain optimal inference. This also ensures that existing methods for maximum likelihood estimation do not break.

We would like to note again, in light of Corollary 1, that in general, using noisy sufficient statistics z of any model instead of the true sufficient statistics may lead to inconsistent estimates, in particular, nonexistence of MLE. A key issue is that the noisy statistic z usually lies in \mathbb{R}^n whereas the validity of many inference procedures (such as existence of MLE and consistency) is guaranteed only when z lies in some set $S \subset \mathbb{R}^n$, typically the convex hull of sufficient statistics of the associated model, for example, $S = \text{Conv}(DS_n)$. In some cases, z is post-processed and projected onto a set S' ; the choice of S' is motivated with a goal of imposing some reasonable constraint on the noisy statistic, and to reduce the L_2 error between the noisy and the original statistics. But usually, $S \neq S'$. We showed with the degree partition example that such approach does not even guarantee the existence of MLE, let alone consistency. Thus more carefully designed and provable methods are needed to guarantee utility, keeping in mind the end

goals of statistical inference (e.g., estimation of parameters, and not just statistics).

We demonstrated that significant gains in utility can be made by using a two step technique of (a) “de-noising” the noisy statistic using maximum likelihood estimation on the measurement error model and (b) estimating the MLE of the parameter of interest using the de-noised version of the statistic. Note that the first step is equivalent to “projecting” the noisy statistic onto the lattice points of the corresponding marginal polytope. While this two step procedure guarantees that the MLE of the parameter exists, a priori, there is no reason to believe that the estimates are also consistent and asymptotically normal. But we prove, remarkably, in the case of β -model, that they are. We believe that this principled two step approach could be applicable in other settings, and would lead to not only existence of MLE but also consistency and asymptotic normality. An interesting class of models to extend these techniques to are the general class of discrete exponential families and in particular, various families of β -models such as the Rasch models of bipartite graphs [e.g., Rinaldo, Petrović and Fienberg (2013)], models based on weighted degree sequences such as those studied in Hillar and Wibisono (2013) and degree sequences of directed graphs, and finally the class of log-linear models where Fienberg, Rinaldo and Yang (2010) have already demonstrated some of the above mentioned issues with estimations done in a privacy-preserving manner.

There are several challenges in extending our principles to the above mentioned class of models. One of the key challenges is, for each of these families, finding a description of the marginal polytope S that would allow the “de-noising” step; the marginal polytope is a complex combinatorial object associated with the existence of MLE and is a focus of many studies; see, for example, Rinaldo, Fienberg and Zhou (2009), but its characterization is often nontrivial. One avenue for further work is to use the technique of asymmetrization of a polytope, as done in this paper, to derive efficient conditions for the existence of MLE for generalized β -models. Once such a description is found, the next challenge is to devise an efficient algorithm for “projecting” the noisy statistic onto the set of lattice points of the marginal polytope. The projection can be informed by the measurement error model. In our case, the significant contribution is achieved, by combing these two steps—finding the “right” description of S and a projection algorithm—into one step. We do this by using an efficient algorithmic description of the lattice points of the marginal polytopes (e.g., the Havel–Hakimi algorithm [Havel (1955), Hakimi (1962)] provides such a description for degree sequences) and somewhat surprisingly, converting such a description into an efficient projection algorithm. Such efficient descriptions do not exist for the more general class of discrete exponential families [e.g., see Hillar and

Wibisono (2013) and Engström and Norén (2010)] and is an interesting direction of future work that goes beyond private estimation and warrants an independent inquiry.

In cases where de-noising is not possible, for example with more general graph statistics, how can we capture the noise infusion due to privacy or some other mechanism? An alternative is to develop new statistical procedures that integrate the noise addition process into the likelihood by using missing data techniques, for example, see Karwa, Slavković and Krivitsky (2014) for differentially private estimation of exponential random graph models. But such solutions may be computationally expensive and currently lack theoretical properties.

8. Proofs.

8.1. *Proof of Theorem 1.* The key technique to prove this result is to use the polytope of degree partitions to characterize the boundary of the polytope of degree sequence, $\text{Conv}(DS_n)$. We will need the following result from Mahadev and Peled (1995) that characterizes the boundary of $\text{Conv}(DS_n)$.

LEMMA 4 [Lemma 3.3.13 in Mahadev and Peled (1995)]. *Let d be a degree sequence of a graph G that lies on the boundary of $\text{Conv}(DS_n)$. Then there exist nonempty and disjoint subsets S and T of $\{1, \dots, n\}$ such that:*

1. S is clique of G ;
2. T is a stable set of G ;
3. Every vertex in S is adjacent to every vertex in $(S \cup T)^c$ in G ;
4. No vertex of T is adjacent to any vertex of $(S \cup T)^c$ in G .

Part (i)—MLE of the degree partition β -model: By Theorem 9.13 in Barndorff-Nielsen (1978), the MLE $\hat{\beta}(\bar{d})$ exists iff $\bar{d} \in \text{ri}(\text{Conv}(DP_n))$. Here, $\text{ri}(\text{Conv}(A))$ denotes the relative interior of the convex hull of A . To prove the first part of the theorem, note that the following system of inequalities along with the constraint $d_1 \leq d_2 \leq \dots \leq d_n$ describe the faces the convex hull of degree partitions [see Theorem 1.3 in Bhattacharya, Sivasubramanian and Srinivasan (2006)]:

1.

$$\bar{d}_i > 0 \quad \text{and} \quad \bar{d}_i < n - 1 \quad \forall i \quad \text{and,}$$

2.

$$\sum_{i=1}^k \bar{d}_i - \sum_{i=n-l+1}^n \bar{d}_i < k(n-1-l) \quad \text{for } 1 \leq k+1 \leq n.$$

Thus, the ordering constraints also define $n - 1$ faces of the polytope given by $d_{i+1} - d_i \leq 0$. For a degree partition to be in the interior of $\text{Conv}(DP_n)$, it must hold that $\bar{d}_1 > \bar{d}_2 > \dots > \bar{d}_n$. This is possible only if each $\bar{d}_i = n - i$. However, such a sequence is not realizable (and hence not a degree sequence) as $\bar{d}_n = 0$ and $\bar{d}_1 = n - 1$. Hence, there is no degree partition that lies in the interior of $\text{Conv}(DP_n)$, and the MLE for the degree partition β -model never exists when we observe only one graph.

Part (ii): We have to show that if the MLE $\hat{\beta}(d)$ exists, then \bar{d} satisfies the system (2.2). Recall that the MLE $\hat{\beta}(d)$ exists iff $d \in \text{ri}(\text{Conv}(DS_n))$. Also, note that $d \in \text{ri}(\text{Conv}(DS_n))$ iff

$$(8.1) \quad \sum_{i \in S} d_i - \sum_{i \in T} d_i < |S|(n-1-|T|) \quad \forall S, T \subset [n], S \cup T \neq \emptyset, S \cap T = \emptyset.$$

For example, see Theorem 3.3.17 in Mahadev and Peled (1995).

We show that the system of inequalities in (8.1) are permutation invariant, that is, if d satisfies (8.1), then πd also satisfies (8.1), where π is any permutation on $[n] = \{1, \dots, n\}$. To see this, let $(\mathcal{S}, \mathcal{T}) = \{(S, T)\}$ be the set of all possible sets S and T such that $S, T \subset [n] = \{1, \dots, n\}$, $S \cup T \neq \emptyset$, $S \cap T = \emptyset$. First, note that if $(S, T) \in (\mathcal{S}, \mathcal{T})$, then $(T, S) \in (\mathcal{S}, \mathcal{T})$. Also, note that $(\mathcal{S}, \mathcal{T})$ is closed under permutations, that is, if $(S, T) \in (\mathcal{S}, \mathcal{T})$, and if π is any permutation on $[n]$, then $(\pi S, \pi T) \in (\mathcal{S}, \mathcal{T})$.

Now assume that $d \in \text{ri}(\text{Conv}(DS_n))$, we need to show that \bar{d} satisfies the system of inequalities (2.2). Note that d satisfies (8.1). By the fact that these inequalities are permutation invariant, any permutation of d also satisfies (8.1). Hence, as $\bar{d} = \pi d$ for some permutation π , (8.1) is true for \bar{d} .

Taking $S = \{1, \dots, k\}$ and $T = \{n - l + 1, \dots, n\}$ gives the second set of inequalities in (2.2). Taking $S = \{i\}$, $T = \emptyset$ gives $\bar{d}_i < n - 1$ and taking $S = \emptyset$, $T = \{i\}$ gives $\bar{d}_i > 0$.

Part (iii): Assume that \bar{d} satisfies the system (2.2). We will show that \bar{d} does not lie on the boundary of $\text{Conv}(DS_n)$. This will imply that $\bar{d} \in \text{ri}(\text{Conv}(DS_n))$, which implies that \bar{d} satisfies the inequalities (8.1). By the permutation invariance of the system (8.1), $\pi \bar{d} = d$ also satisfies (8.1), from which the result follows.

All that is remaining to be shown is that \bar{d} does not lie on the boundary of $\text{Conv}(DS_n)$. The boundary of $\text{Conv}(DS_n)$ is characterized by Lemma 4. Let G be a graph that realizes \bar{d} , hence G is such that there exist disjoint subsets of $\{1, \dots, n\}$ S and T satisfying conditions of Lemma 4.

Let $i \in S$, then $\bar{d}_i \geq (|S| - 1) + |(S \cup T)^c| = n - |T| - 1$ (by conditions 1 and 3 of Lemma 4). Let $i \in T$ then $\bar{d}_i \leq |S|$. Finally if $i \in (S \cup T)^c$, then $\bar{d}_i \geq |S|$ (by condition 3 in Lemma 4) and $\bar{d}_i \leq |S| + |(S \cup T)^c| - 1 = n - |T| - 1$ (by condition 4 in Lemma 4). Putting these together, we get the following:

$$0 \leq \bar{d}_i \leq |S|, \quad i \in T,$$

$$(8.2) \quad \begin{aligned} |S| \leq \bar{d}_i \leq n - |T| - 1, & \quad i \in (S \cup T)^c, \\ n - |T| - 1 \leq \bar{d}_i \leq n - 1, & \quad i \in S. \end{aligned}$$

Now note that $\bar{d}_1 \leq \bar{d}_2 \leq \dots \leq \bar{d}_n$. Hence, the only possible choice for S and T are $S = \{1, \dots, k\}$ and $T = \{n - l + 1, \dots, n\}$ where $k = |S|$, $l = |T|$, $1 \leq k + l \leq n$. No other combinations of S and T exist, due to the characterization of \bar{d} given in equation (8.2). Next, since \bar{d} is on the boundary of $\text{Conv}(DS_n)$, it holds that $\sum_{i \in S} d_i - \sum_{i \in T} d_i = |S|(n - 1 - |T|)$ for all such S and T described above. However, we are given that this is not true. Hence \bar{d} must lie in the interior of $\text{Conv}(DS_n)$.

Acknowledgments. Karwa was a graduate student at the Department of Statistics, Pennsylvania State University when the paper was initially submitted. The authors would also like to thank anonymous referees, editor, Alessandro Rinaldo and Johannes Rau for helpful feedback.

SUPPLEMENTARY MATERIAL

Supplement to “Inference using noisy degrees: Differentially Private β -model and synthetic graphs” (DOI: [10.1214/15-AOS1358SUPP](https://doi.org/10.1214/15-AOS1358SUPP); .pdf). This supplementary material contains the proof of the key Theorems 2, 3 and 4 from the paper.

REFERENCES

- ARRATIA, R. and LIGGETT, T. M. (2005). How likely is an i.i.d. degree sequence to be graphical? *Ann. Appl. Probab.* **15** 652–670. [MR2114985](#)
- BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester. [MR0489333](#)
- BHATTACHARYA, A., SIVASUBRAMANIAN, S. and SRINIVASAN, M. K. (2006). The polytope of degree partitions. *Electron. J. Combin.* **13** Research Paper 46, 18 pp. (electronic). [MR2223521](#)
- BLITZSTEIN, J. and DIACONIS, P. (2010). A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Math.* **6** 489–522. [MR2809836](#)
- CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. Chapman & Hall/CRC, Boca Raton, FL. [MR2243417](#)
- CHATTERJEE, S., DIACONIS, P. and SLY, A. (2011). Random graphs with a given degree sequence. *Ann. Appl. Probab.* **21** 1400–1435. [MR2857452](#)
- DUCHI, J. C., JORDAN, M. I. and WAINWRIGHT, M. J. (2013). Local privacy, data processing inequalities, and statistical minimax rates. Preprint. Available at [arXiv:1302.3203](https://arxiv.org/abs/1302.3203).

- DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006a). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography. Lecture Notes in Computer Science* **3876** 265–284. Springer, Berlin. [MR2241676](#)
- DWORK, C., KENTHAPADI, K., MCSHERRY, F., MIRONOV, I. and NAOR, M. (2006b). Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology—EUROCRYPT 2006. Lecture Notes in Computer Science* **4004** 486–503. Springer, Berlin. [MR2423560](#)
- ENGSTRÖM, A. and NORÉN, P. (2010). Polytopes from subgraph statistics. Preprint. Available at [arXiv:1011.3552](#).
- 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 SLAVKOVIĆ, A. B. (2010). Data privacy and confidentiality. In *International Encyclopedia of Statistical Science* 342–345. Springer, Berlin.
- GHOSH, A., ROUGHGARDEN, T. and SUNDARARAJAN, M. (2009). Universally utility-maximizing privacy mechanisms. In *STOC'09—Proceedings of the 2009 ACM International Symposium on Theory of Computing* 351–359. ACM, New York. [MR2780081](#)
- GOODREAU, S. M., KITTS, J. A. and MORRIS, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* **46** 103–125.
- HAKIMI, S. L. (1962). On realizability of a set of integers as degrees of the vertices of a linear graph. I. *J. Soc. Indust. Appl. Math.* **10** 496–506. [MR0148049](#)
- HANDCOCK, M. S. and GILE, K. J. (2010). Modeling social networks from sampled data. *Ann. Appl. Stat.* **4** 5–25. [MR2758082](#)
- HAVEL, V. (1955). A remark on the existence of finite graphs. *Casopis Pest. Mat.* **80** 477–480.
- HAY, M., LI, C., MIKLAU, G. and JENSEN, D. (2009). Accurate estimation of the degree distribution of private networks. In *Ninth IEEE International Conference on Data Mining, ICDM'09* 169–178. IEEE, New York.
- HELLERINGER, S. and KOHLER, H.-P. (2007). Sexual network structure and the spread of HIV in Africa: Evidence from Likoma island, Malawi. *AIDS* **21** 2323–2332.
- HELLERINGER, S., KOHLER, H.-P., CHIMBIRI, A., CHATONDA, P. and MKANDAWIRE, J. (2009). The Likoma network study: Context, data collection, and initial results. *Demogr. Res.* **21** 427–468.
- HILLAR, C. and WIBISONO, A. (2013). Maximum entropy distributions on graphs. Preprint. Available at [arXiv:1301.3321](#).
- HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. [MR0608176](#)
- HUNDEPOOL, A., DOMINGO-FERRER, J., FRANCONI, L., GIESSING, S., NORDHOLT, E. S., SPICER, K. and DE WOLF, P.-P. (2012). *Statistical Disclosure Control*. Wiley, Chichester. [MR3026260](#)
- HUNTER, D. R. (2004). MM algorithms for generalized Bradley–Terry models. *Ann. Statist.* **32** 384–406. [MR2051012](#)
- HUNTER, D. R., GOODREAU, S. M. and HANDCOCK, M. S. (2008). Goodness of fit of social network models. *J. Amer. Statist. Assoc.* **103** 248–258. [MR2394635](#)
- KARWA, V. and SLAVKOVIĆ, A. B. (2012). Differentially private graphical degree sequences and synthetic graphs. In *Privacy in Statistical Databases* 273–285. Springer, Berlin.
- KARWA, V. and SLAVKOVIĆ, A. (2015). Supplement to “Inference using noisy degrees: Differentially private β -model and synthetic graphs.” DOI:[10.1214/15-AOS1358SUPP](#).

- KARWA, V., SLAVKOVIĆ, A. B. and KRIVITSKY, P. (2014). Differentially private exponential random graphs. In *Privacy in Statistical Databases* 143–155. Springer, Berlin.
- KASIVISWANATHAN, S. P., NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2013). Analyzing graphs with node differential privacy. In *Theory of Cryptography* 457–476. Springer, Berlin.
- MAHADEV, N. V. and PELED, U. N. (1995). *Threshold Graphs and Related Topics*. Elsevier, Amsterdam.
- NARAYANAN, A. and SHMATIKOV, V. (2009). De-anonymizing social networks. In *30th IEEE Symposium on Security and Privacy* 173–187. IEEE, New York.
- NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2007). Smooth sensitivity and sampling in private data analysis. In *STOC'07—Proceedings of the 39th Annual ACM Symposium on Theory of Computing* 75–84. ACM, New York. [MR2402430](#)
- OGAWA, M., HARA, H. and TAKEMURA, A. (2011). Graver basis for an undirected graph and its application to testing the beta model of random graphs. Preprint. Available at [arXiv:1102.2583](#).
- OLHEDE, S. C. and WOLFE, P. J. (2012). Degree-based network models. Preprint. Available at [arXiv:1211.6537](#).
- PERRY, P. O. and WOLFE, P. J. (2012). Null models for network data. Preprint. Available at [arXiv:1201.5871](#).
- RAMANAYAKE, A. and ZAYATZ, L. (2010). Balancing disclosure risk with data quality. Statistical Research Division Research Report Series No. 2010-04, U.S. Census Bureau, Washington, DC.
- RINALDO, A., FIENBERG, S. E. and ZHOU, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.* **3** 446–484. [MR2507456](#)
- RINALDO, A., PETROVIĆ, S. and FIENBERG, S. E. (2013). Maximum likelihood estimation in the β -model. *Ann. Statist.* **41** 1085–1110. [MR3113804](#)
- SADEGHI, K. and RINALDO, A. (2014). Statistical models for degree distributions of networks. Preprint. Available at [arXiv:1411.3825](#).
- SAMPSON, S. F. (1968). A novitiate in a period of change: An experimental and case study of social relationships Ph.D. thesis, Cornell Univ., Ithaca, NY.
- SMITH, A. (2008). Efficient, differentially private point estimators. Preprint. Available at [arXiv:0809.4794](#).
- SNIJDERS, T. A. B. (2003). Accounting for degree distributions in empirical analysis of network dynamics. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers* 146–161. The National Academies Press, Washington, DC.
- VU, D. and SLAVKOVIĆ, A. (2009). Differential privacy for clinical trial data: Preliminary evaluations. In *IEEE International Conference on Data Mining Workshops, ICDMW'09* 138–143. IEEE, New York.
- WASSERMAN, L. and ZHOU, S. (2010). A statistical framework for differential privacy. *J. Amer. Statist. Assoc.* **105** 375–389. [MR2656057](#)
- WILLENBORG, L. and DE WAAL, T. (1996). *Statistical Disclosure Control in Practice*. Springer, New York.
- 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. [MR3068452](#)
- YAN, T., ZHAO, Y. and QIN, H. (2015). Asymptotic normality in the maximum entropy models on graphs with an increasing number of parameters. *J. Multivariate Anal.* **133** 61–76. [MR3282018](#)
- ZACHARY, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* **33** 452–473.

ZHANG, J. and CHEN, Y. (2013). Sampling for conditional inference on network data. *J. Amer. Statist. Assoc.* **108** 1295–1307. [MR3174709](#)

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
132G BAKER HALL
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: vishesh@cmu.edu

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
421A THOMAS BLDG.
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: sesa@psu.edu