SAND2020-5256C

Differentially Private Generation of Social Networks via Exponential Random Graph Models

Fang Liu[1], Evercita C. Eugenio[2], Ick Hoon Jin[3], Claire Bowen[5]

[1]*Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46530*
[2]*Data Science and Cyber Analytics, Sandia National Laboratories, Livermore, CA 94550*
[3]*Yonsei University*
[4]*Urban Institute*
fliu2@nd.edu, eceugen@sandia.gov, ijin@yonsei.ac.kr, cbowen@urban.org

***Abstract*—Many social networks contain sensitive relational information. One approach to protect the sensitive relational information while offering flexibility for social network research and analysis is to release synthetic social networks at a pre-specified privacy risk level, given the original observed network. We propose the DP-ERGM procedure that synthesizes networks that satisfy the differential privacy (DP) via the exponential random graph model (EGRM). We apply DP-ERGM to a college student friendship network and compare its original network information preservation in the generated private networks with two other approaches: differentially private DyadWise Randomized Response (DWRR) and Sanitization of the Conditional probability of Edge given Attribute classes (SCEA). The results suggest that DP-EGRM preserves the original information significantly better than DWRR and SCEA in both network statistics and inferences from ERGMs and latent space models. In addition, DP-ERGM satisfies the node DP, a stronger notion of privacy than the edge DP that DWRR and SCEA satisfy.**

***Index Terms*—exponential random graph model (ERGM), goodness of fit, node differential privacy (DP), social networks, Bayesian, posterior distribution.**

I. Introduction

For the last few decades, social network (SN) analysis and research have grown tremendously, especially with the emergence of social media (e.g., Facebook and Twitter). While the voluminousness and popularity of social network data have enabled new discoveries, they have also increased privacy risk of individuals. For instance, the Cambridge Analytica leveraged on the "friendship" function in Facebook and landed the personal data of multi-millions of Facebook users' profiles for political advertising purposes [1]. Network data of sexual relationships and sexually transmitted diseases are extremely sensitive information, disclosure of which can lead to social reactions and stigma that can negatively affect the individuals in the study [2, 3].

The state-of-art research work on protecting the privacy of graph or network data is largely built upon the concept of differential privacy (DP) [4], which provides a rigorous mathematical guarantee on privacy protection. Two notions of DP for relational data have been proposed: edge DP and node DP. A procedure that satisfies the edge DP ensures that its output does not reveal more information regarding a particular relation on top of what the data intruder already knows, while the output from a procedure of node DP does not reveal more information regarding the relationships between a particular node with the rest of the nodes in a network on top of what the data intruder already knows. Therefore, the node DP considers provides a stronger guarantee of privacy protection than edge DP and is more relevant for preserving the global structure of a network, which is the focus of this paper. There are exiting approaches on protecting network privacy in the setting of node DP. For example, [5] apply the Johnson-Linderstrauss transform to release the number of edges crossed in a graph cut; [6] develop techniques that project the original graph onto the set of graphs with the maximum degree below a certain threshold but restrict the utility analysis to linear functions of the degree distribution or subgraph counting; [7] introduce mechanisms for private sub-graph counting. [8] develop Lipschitz extensions and generalize the exponential mechanism [9] (a DP mechanism for releasing numerical and non-numerical queries) to approximate the degree distribution.

In summary, existing approaches in the framework of node DP focus on releasing specific graph summary statistics. For social network data analysis, merely outputting a limited set of summary statistics of a network might not be satisfactory from the practical or research perspective. One solution to accommodate the practical and research needs on network data while ensuring the privacy of the networks in the DP framework is to release differentially private surrogate or synthetic networks, so that users can perform their own analysis as if they had the original network data.

There exist a few approaches for differentially private synthesis of network data and all of them satisfy the weaker edge DP. [10] introduce an algorithm that releases synthetic relational data based on differentially private β models. β models are one simplest type of the exponential random graph model family. Simplicity is also the biggest disadvantage of this approach in that synthetic networks may greatly deviate from the observed network. [11] develop a privacy preserving network generator based on dK-graph models [12]. dK graph models consider degree correlations among $d \geq 0$ nodes.

Fang Liu and Evercita Eugenio are co- first authors. Fang Liu and Evercita Eugenio were supported by the National Science Foundation (NSF) Grants #1546373 and #1717417, Ick Hoon Jin was partially supported by the Yonsei University Research Fund of 2019-22-0210 and National Research Foundation of Korea (NRF 2020R1A2C1A01009881), and Claire McKay Bowen was supported by the NSF Graduate Research Fellowship under Grant No. DGE-1313583 during part of the paper's development. The publication has been assigned the Sandia National Laboratories identifier XXXXXXXX.

There are two limitations for this approach. First, the privacy budget, which is the pre-set privacy risk tolerance level, needs to be relatively large to have some utility in generated graphs, where utility refers to preservation of key original network information. Second, algorithms for generating graphs when $d \geq 3$ do not exist. [13] develop a differentially private dyad-wise randomized response (DWRR) approach. the DWRR is straightforward and easy to implement, but synthetic networks tend to be very dense unless the privacy budget is large. In addition, each edge is perturbed locally, which could distort the global structure of the original network. Finally, DWRR assigns a separate privacy budget when sanitizing each edge and does not quantify the total privacy cost from releasing the whole network. If different edges are not independent, the total privacy cost from the whole network will exceed the nominal per-edge privacy budget. In summary, all the above mentioned private network synthesis approaches have their respective drawbacks in either utility or DP. In addition, acceptable utility of synthetic networks is only observed for relatively high privacy budgets.

Methods for synthesizing networks that guarantee the privacy of the released whole network while maintaining its utility are still in great need. We propose a new approach, DP-ERGM, that generates private networks via exponential random graph models (ERGMs). We choose to use ERGMs because they are flexible generative models for networks, incorporating important topological and nodal information, and can model complex relationships among nodes. Compared to the existing approaches for private network synthesis, DP-ERGM has the following advantages. First, from a privacy protection perspective, it satisfies the node DP for the whole network and thus provides stronger guarantee of privacy protection. Second, the synthetic networks via DP-ERGM are expected to have higher utility for a large range of network structures. In contrast, DP-ERGM is based on the general ERGM framework.

II. Methodology

A. Preliminaries on Differential Privacy (DP)

Definition 1 (**DP** [4])**.** A sanitization algorithm $\mathcal{R}$ releases statistics $\mathbf{s}$ with ϵ-DP if for all possible data set pairs $(\mathbf{x}, \mathbf{x}')$ that differ by one record, and all results $Q \subseteq \mathcal{T}$,

$$\left| \log \left(\frac{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x})) \in Q)}{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}')) \in Q)} \right) \right| \leq \epsilon, \tag{1}$$

where $\mathcal{T}$ denotes the output range of $\mathcal{R}(\mathbf{s})$, and $\epsilon > 0$ is the privacy budget.

DP provides a mathematically rigorous framework for protecting individuals when releasing statistics from a data set, regardless of the knowledge and behaviors of data intruders. The privacy budget ϵ is pre-specified and represent the privacy risk for releasing a query from the sanitization algorithm. $\mathcal{R}$. The smaller ϵ is, the smaller the probability of identifying an individual based on the released sanitized query.

There are several commonly used mechanisms that releases statistics ϵ-DP. The Laplace mechanism [4] adds noise to original statistics $\mathbf{s} = (s_1, \ldots, s_K)$ to generate sanitized $\mathbf{s}^* = \mathbf{s} + \mathbf{e}$, where $\mathbf{e} \overset{\text{ind}}{\sim} \text{Laplace}(0, \delta_1/\epsilon)$ and $\delta_1 = \max_{\mathbf{x},\mathbf{x}'} \|\mathbf{s}(\mathbf{x}) - \mathbf{s}(\mathbf{x}')\|_1$ is the l_1 global sensitivity of $\mathbf{s}$, across all paired data sets $\mathbf{x}$ and $\mathbf{x}'$ that differ by one record. The exponential mechanism releases query result s^* with probability $\exp(u(s^*)\epsilon/(2\delta_u))/\sum_{s' \in \mathcal{S}} \exp(u(s')\epsilon/(2\delta_u))$, where δ_u is the maximum change in an utility function u with one element change in data $\mathbf{x}$. Other mechanisms include the Gaussian mechanism [14, 15] that relies on the relaxed DP concepts such as approximate DP [16] and probabilistic DP [17]).

B. Exponential Random Graph Models (ERGMs)

ERGMs are a family of popular statistical models for analyzing social network data [18, 19]. ERGMs help to explain the structure of a network, and support statistical inference of the processes that influence the formation of network structures. They are also effective generative models for network data, accommodating various types of structural dependencies among the nodes in a network. A ERGM is specified as

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}^T \mathbf{S}(\mathbf{y}, \mathbf{x})\right\}/K(\boldsymbol{\theta}), \tag{2}$$

where $\mathbf{y}$ is the $n \times n$ adjacency matrix among the nodes in a social network ($y_{ij} = 1$ if an edge exists between node i and node j, $y_{ij} = 0$ otherwise); $\mathbf{x}$ is $n \times q$ matrix that contains q nodal attributes of the n nodes; and $\mathbf{S}(\mathbf{y}, \mathbf{x})$ is a column vector that contains summary statistics of the network and often include metrics on the network structure as well as nodal statistics that might relate to $\mathbf{y}$; and $\boldsymbol{\theta}$ is of the same dimension as $\mathbf{S}(\mathbf{y}, \mathbf{x})$ and contains the model parameters and represents the effects of $\mathbf{S}(\mathbf{y}, \mathbf{x})$ on the network structure. $K(\boldsymbol{\theta})$ in Eqn (2) is an analytically intractable normalizing constant summed over all possible adjacency matrix $\mathbf{y}'$.

C. Differentially Private Social Network Synthesis via ERGM (DP-ERGM)

Leveraging the properties and functionalities of ERGMs, we propose the DP-ERGM procedure to synthesize differentially private networks. DP-ERGM is based on a Bayesian framework. The steps of the DP-ERGM procedure are provided in Algorithm 1. In brief, we first obtain the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x})$ given the likelihood function in Eqn (2) and a data user-specified prior on $\boldsymbol{\theta}$. We then sanitize $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x})$ to obtain differentially private $p^*(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x})$ and draw a sample of $\boldsymbol{\theta}^*$, which will be used by the ERGM model in Eqn (2) to generate a network. The process of sanitization and drawing can be repeated $m > 1$ times to generate m synthetic networks so to capture the sanitization and the synthesis uncertainty to allow for valid statistical inferences on $\boldsymbol{\theta}$ given the released networks. The inferences for a statistical model fitted to the multiple sets can be combined with the formulas given in [20] and [21]. To preserve DP in releasing m sets of networks, each synthetic set is allocated a privacy budget of $1/m$ of the total privacy budget ϵ per the sequential composition. Our empirical studies suggests $m = 3$ to 5 is a good choice in general and provides enough information across multiple synthetic sets to capture the sanitization and synthesis uncertainty without injecting too much DP noise in each individual synthetic set.

Algorithm 1 DP-ERGM

1: **Input**: original network $(\mathbf{x}, \mathbf{y})$, a prespecified EGRM $\mathcal{M}$ or a set of candidate EGRMs, privacy budget ϵ; number of synthetic networks m.
2: **Output**: m differentially private synthetic networks
3: **Do** $k = 1, \ldots, m$
4: If a set of ERGMs are given, select a model $\mathcal{M}^{(k)}$ via the exponential mechanism with privacy budget ϵ'/m, where ϵ' is the allocated budget for model selection.
5: Run model $\mathcal{M}^{(k)}$ on $(\mathbf{x}, \mathbf{y})$ in the Bayesian framework to obtain the empirical or an approximate posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$.
6: Obtain differentially private posterior sample $\boldsymbol{\theta}^{*(k)}$ with budget $(\epsilon - \epsilon')/m$.
7: Feed $\boldsymbol{\theta}^{*(k)}$, nodal information $\mathbf{x}$, and $\mathcal{M}^{(k)}$ to Algorithm 2 to generate a differentially private network.
8: **End Do**

Algorithm 2 Generation of networks from ERGM via Monte Carlo Markov chain sampling

1: **Input**: ERGM($\boldsymbol{\theta}, \mathbf{x}$); MCMC iterations T.
2: **Output**: a random network sample from ERGM($\boldsymbol{\theta}, \mathbf{x}$)
3: Initialize a network $\mathbf{y}^{(0)}$. Calculate statistics $\mathbf{S}^{(0)}$ associated with ERGM($\boldsymbol{\theta}, \mathbf{x}$).
4: **Do** $t = 1, \ldots, T$
5: Randomly choose a pair of nodes (i, j) from $\mathbf{y}^{(t-1)}$ and flip the edge between them to propose a candidate network $\mathbf{y}^c$.
6: Calculate summary statistics $\mathbf{S}^c$ given $(\mathbf{x}, \mathbf{y}^c)$.
7: Set $\mathbf{y}^{(t)} = \mathbf{y}^c$ with probability $\min(1, \pi)$, where $\pi = \exp\left(\boldsymbol{\theta}^T(\mathbf{S}^c - \mathbf{S}^{(t-1)})\right)$.
8: **End Do**

Following the current practice in generating differentially private synthetic networks, Algorithm 1 focuses on synthesizing relations/edges $\mathbf{y}$, and the nodes in the sanitized networks are left as the original. If there are nodal attributes in the original network, which are deemed sensitive in addition to the relations, one may allocate a portion of the total budget ϵ to generate differentially private nodes $\mathbf{x}^{*(k)}$. Since nodal data are often presented in tabular forms, there exist various approaches for sanitizing this type of data, such as the Laplace sanitizer of the full-dimensional histogram of $\mathbf{x}$. Readers may refer to [21] for a brief review on differentially private synthesis of tabular data.

Proposition 1. The synthetic network via the DP-ERGM procedure in Algorithm 1 satisfies the node-DP for a given ERGM.

The detailed proof of the proposition will be provided in an extended paper. Briefly, the predictive posterior distribution $p(\mathbf{x}^*, \mathbf{y}^*|\mathbf{x}, \mathbf{y})$ can be written as $\int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) p(\mathbf{x}^*|\mathbf{x}) d\boldsymbol{\theta}$. $p(\boldsymbol{\theta}^*|\mathbf{x}, \mathbf{y})$ satisfies the node DP per the nature of ERGMs. The differentially private sanitization of $p(\mathbf{x}^*|\mathbf{x})$ does involve edges, so strictly speaking, the differentiation between the node DP vs. edge DP really does not apply to $p(\mathbf{x}^*|\mathbf{x})$. Denote by ϵ_x the budget allocated to sanitize $\mathbf{x}$ and $\mathcal{G} = (\mathbf{x}, \mathbf{y})$ and $\mathcal{G}' = (\mathbf{x}', \mathbf{y}')$ two networks differing by one node and the corresponding changes in its relations with other nodes, then

$$\begin{aligned}
&\frac{p(\mathbf{x}^*, \mathbf{y}^*|\mathbf{x}, \mathbf{y})}{p(\mathbf{x}^*, \mathbf{y}^*|\mathbf{x}', \mathbf{y}')} = \frac{p(\mathbf{x}^*|\mathbf{x}) \int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) d\boldsymbol{\theta}}{p(\mathbf{x}^*|\mathbf{x}') \int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}', \mathbf{y}') d\boldsymbol{\theta}} \\
\leq & e^{\epsilon_x} \frac{\int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) d\boldsymbol{\theta}}{\int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}', \mathbf{y}') d\boldsymbol{\theta}} \\
\leq & e^{\epsilon_x} \frac{\int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta}) e^{\epsilon - \epsilon_x} p(\boldsymbol{\theta}|\mathbf{x}', \mathbf{y}') d\boldsymbol{\theta}}{\int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}', \mathbf{y}') d\boldsymbol{\theta}} = e^{\epsilon}.
\end{aligned}$$

III. Application

To evaluate the statistical and inferential utility of the synthetic network data generated by DP-ERGM, we apply DP-ERGM to the Chinese college student friendship network and benchmark its performance against some existing private edge synthesis approaches. The Chinese college student friendship network contains 162 students from a four-year college in China collected by the Lab for Big Data Methodology in the Department of Psychology at the University of Notre Dame in 2017 [22]. There are 848 edges among the 162 students, representing friendship (Fig 1). The nodal attributes include the students' gender, grade point average (GPA), class, and the number of cigarettes smoked per day in the past 30 days.

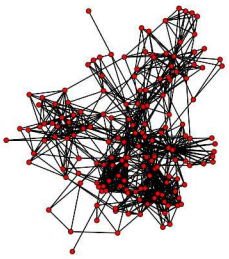

Fig. 1: Chinese college student friendship network

A. Synthesis Procedures

We examine three privacy budget settings $\epsilon \in (e^{-1}, e, e^2)$, presenting small, moderate, and relatively privacy risk. We generate $m = 4$ differentially private networks at each ϵ. To examine the stability of the synthesis methods, 100 repetitions were run for the college friendship example at each ϵ.

For DP-ERGM, we applied Algorithm 1 to generate 4 networks with synthetic edges from the ERGM in Eqn (2) with $\mathbf{S}(\mathbf{y}, \mathbf{x}) = \{\text{GWD}(\mathbf{y}),\ \text{GWESP}(\mathbf{y}),\ |\text{GPA}_i\text{-GPA}_j|,\ |\#\ \text{cigarettes}_i - \#\ \text{cigarettes}_j|,\ 1(\text{class}_i = \text{class}_j),\ 1(\text{gender}_i = \text{gender}_j)\}$. i and j are indices for nodes; 1 is an indicator function; GWD stands for geometrically weighted degree and is defined as $e^{\tau} \sum_{i=1}^{n-1} \{1 - (1 - e^{-\tau})^i\} \text{D}_i$ for $\tau > 0$ and D_i for $i = 0, 1, \cdots, n-1$ represent the number of nodes whose degree is i with constraint $\sum_{i=0}^{n-1} \text{D}_i = n$; and GWESP stands for geometrically weighted edgewise shared partnership (ESP) and is defined as $e^{\tau'} \sum_{k=1}^{n-2} \{1 - (1 - e^{-\tau'})^i\} \text{ESP}_i$ for $\tau' > 0$ and ESP_i for $i = 1, 2, \cdots, n-2$ represents the number of edges whose endpoints both share edges with exactly i other nodes with constraint $\sum_{i=0}^{n-2} \text{ESP}_i =$ the number of edges in the network.

In addition, the GOF statistics in Figure 2 suggests the mode without # of edges as a covariate captures the original network information well. we assumed a non-informative prior $f(\boldsymbol{\theta}) \propto$ const and leveraged the asymptotic normality assumption of the posterior distribution of $\boldsymbol{\theta}$ to draw and sanitize the posterior samples.

[13] propose DWRR for sharing social network data by synthesizing edges via randomized response and suggest $p_{ij} = q_{ij} = 1 - \pi_{ij} = e^{\epsilon_{ij}}/(1+e^{\epsilon_{ij}})$ for all $i \neq j = 1, \ldots, n$, where p_{ij} is the probability of retaining an edge nodes (i, j) and q_{ij} is the probability of retaining the absence of an edge, and ϵ_{ij} is privacy budget for retaining the edge DP between nodes (i, j). In their numerical examples, $\epsilon_{ij} \equiv \epsilon$ and the probability of edge flipping is a constant $\pi = 1/(1+e^{\epsilon})$. In our case, $\pi = 1/(1+e^{\epsilon/4})$ (divided by 4 because 4 networks were generated and released). For the three budgets $\epsilon = e^{-1}, e$ and e^2 is $\pi = 1/(1+e^{\epsilon/4})$, π equals to 0.477, 0.336, and 0.136, respectively.

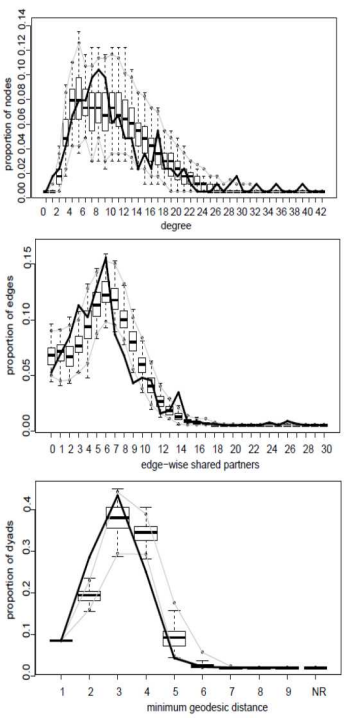

Fig. 2: Goodness of fit of the ERGM in the observed college friendship network

We also designed an intuitive approach that synthesizes the conditional distribution of edges given nodal attributes, referred to as the SCEA approach. SCEA sanitizes the edge probability between two nodes classes, where a node class refers to a cell from the full dimensional cross-tabulation of the nodes attributes. SCEA assumes whether a pair of nodes within the same class are tied or not follows a Bernoulli distribution with a within-class rate $p_{k,k}$; and whether a pair of nodes from two different classes (k, k') are connected is governed by another Bernoulli distribution with a between-class rate $p_{k,k'}$. $p_{k,k}$ makes the diagonal elements and $p_{k,k'}$ makes the off-diagonal elements in the $K \times K$ probability matrix $\mathbf{P}$. For this application, there are 4 nodal covariates: class (6 levels), gender (2 levels), GPA converted to a 5-letter grade and number of cigarettes grouped into 3 groups (0, $[1, 10)$, ≥ 10). The cross-tabulation of the 4 nodal attributes leads to 180 cells. After discarding the empty cells, the final $\mathbf{P}$ matrix is of dimension 57×57. SCEA sanitizes $\mathbf{P}$ via a DP mechanism (such as the Laplace mechanism) to obtain $\mathbf{P}^*$ that satisfies the edge DP. Once the differentially private $\mathbf{P}^*$ is obtained, then synthetic edges between nodes within class k are sampled independently from Bern(p^*_{kk}), and those between nodes from classes k and k' are sampled independently from Bern$(p^*_{kk'})$.

B. Results

Examples of the differentially private synthetic networks are provided in Figures 3. DWRR produces very dense networks at all examined ϵ values. SCAE tends to produce dense networks as well though not to the same degree as DWRR. The synthetic networks by DP-ERGM have a similar level of denseness as the original network.

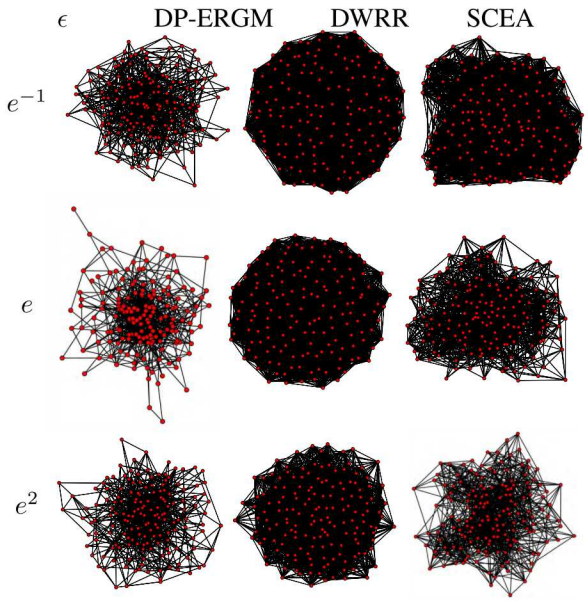

Fig. 3: Examples of differentially private synthetic college student friendship network.

To assess the utility of the differentially private synthetic networks, we obtain the summary statistics of edge counts, GWESP and GWD, and run the ERGM and the latent space model [23] on the synthetic networks from the three synthesis approaches, and compared the private statistics and inferences to the original. In the latent space model, we included covariates of the absolute differences between nodes i and j on GPA and the number cigarettes, and node matching on class and gender. We set the the dimension in the latent space at 2. The average deviations and the root mean squared deviations are presented in Table I for the number of edges, GWESP, and GWD based on the synthetic data from those based on the original data, and in Table II for the ERGM model parameter estimates.

TABLE I: Relation summary statistics from the synthetic college friendship networks.

	original	edges		GWESP			GWD		
privacy	statistic	848		2472.15			398.25		
budget	**DP-ERGM**	DWRR	SCEA	**DP-ERGM**	DWRR	SCEA	**DP-ERGM**	DWRR	SCEA
	average deviation from the original								
e^{-1}	**0.818**	5410	1148	**-590.6**	24203	2404	**-1.291**	20.64	200.4
e	**-1.045**	3818	-51.35	-596.0	17267	**360.6**	**-1.638**	20.64	200.4
e^2	**-0.630**	1542	219.5	**-600.5**	5175	1104	**-1.405**	20.64	200.4
	root mean squared deviation								
e^{-1}	**53.09**	5410	1150	**632.2**	24204	2409	3.406	20.64	200.4
e	**6.545**	3818	56.96	599.4	17268	**367.0**	**3.304**	20.64	200.4
e^2	**2.876**	1542	59.96	604.1	51769	**367.5**	**3.304**	20.64	200.4

DP-ERGM is the obvious winner in both analyses with the smallest deviations on all the examined relations statistics and almost all ERGM parameters at all three ϵ values, except for θ associated with Gender and Class. For GWD in Table I, there is little change over ϵ for DWRR and SCEA on both the deviation and the root mean squared deviation; in addition, the latter appears the same as the former. The similarity between the root mean squared deviation and the average deviation suggests there is little variation in GWD over the 100 repeats of the 4 sets of synthetic networks in the case of DWRR and SCAE, and the main contribution to the root mean squared

TABLE II: ERGM parameter estimates based on the synthetic college friendship networks.

Privacy budget	parameter	θ_{GWESP}	θ_{GWD}	θ_{GPA}	$\theta_{\#\text{cigarettes}}$	θ_{gender}	θ_{class}
	original	-0.922	5.592	0.073	-0.002	0.667	0.887
average deviation from the original							
e^{-1}	**DP-ERGM**	**-0.209**	**-0.480**	**-0.026**	**0.005**	**-0.298**	-0.338
	DWRR	1.271	456.3	-0.087	0.020	-0.638	-0.816
	SCEA	-2.121	-1.076×10^{20}	-0.311	-0.054	0.457	**-0.198**
e	**DP-ERGM**	**-0.213**	**-0.590**	**-0.020**	0.005	-0.298	**-0.334**
	DWRR	-0.943	-79.26	-0.073	**0.000**	-0.595	-0.463
	SCEA	-1.256	-1.580×10^{4}	-0.314	-0.048	**0.009**	-0.916
e^2	**DP-ERGM**	**-0.215**	**-0.519**	**-0.025**	0.006	-0.292	**-0.338**
	DWRR	-0.862	527.6	-0.075	**0.005**	-0.423	0.295
	SCEA	-1.248	-77.68	-0.208	-0.025	**-0.141**	-1.417
root mean squared deviation							
e^{-1}	**DP-ERGM**	**0.211**	**0.924**	**0.048**	**0.007**	0.302	**0.345**
	DWRR	2.802	2.64×10^{4}	0.152	0.151	0.673	0.836
	SCEA	1.365	8.518×10^{20}	0.165	0.028	**0.236**	0.459
e	**DP-ERGM**	**0.214**	**0.881**	**0.042**	0.007	0.301	0.341
	DWRR	0.955	747.3	0.076	**0.002**	0.596	0.464
	SCEA	0.631	5.66×10^{4}	0.162	0.024	**0.041**	**0.116**
e^2	**DP-ERGM**	**0.217**	**0.866**	**0.046**	**0.007**	0.297	**0.346**
	DWRR	0.890	625.9	0.080	0.050	0.423	0.357
	DP-ERGM	0.626	126.7	0.110	0.013	**0.079**	0.709

deviation comes from the deviation in both cases.

The GOF plots from the fitted ERGMs are presented in Figure 4 (due to space limitation, the GOF plots are only presented for $\epsilon = e^{-1}$ and $\epsilon = e^2$. and those at $\epsilon = e^{-1}$ are available in the supplementary materials). The distributions of the GOF statistics from the synthetic networks via DP-ERGM have the best overlap with those based on the original network across all ϵ values, with mild deviation on ESP. For DWRR, the distributions of the GOF statistics deviate significantly from the original for geodesic distance at all ϵ values, and for degree and ESP at small ϵ but improve as ϵ increases. The poor GOF statistics from SCAE suggest the synthetic networks via SCAE preserve poorly the key original information for fitting the EGRM.

The averaged deviation and the root mean squared deviations of the parameter estimates from the latent space model are presented in Table III. Among the four model parameters, DP-EGRM has the smallest average deviations and root mean squared deviations for two parameters, and DWRR and SCEA each has one.

The GOF plots from the latent space model based on the synthetic networks are presented in Figure 5. The GOF statistics from the synthetic networks via DP-ERGM have the best overlap with those based on the original network across all ϵ values with some mild deviation from the original in terms of degree and ESP. For DWRR and SCEA, the degree and geodesic distance measures based on the synthetic networks at all levels of ϵ and the ESP at small ϵ deviate significantly from the original. There is improvement for all three GOF statistics in the case of DWRR as ϵ increases.

IV. Conclusions and Discussion

We propose a new approach DP-ERGM that generates synthetic networks with the node DP. DP-ERGM is based on a generative model that considers high transitivity relations whereas DWRR and SCEA are "local" generative processes that independently sanitize edges, not taking into account the high-order relationships. For that reason, it is not unexpected that the empirical results suggest that DP-ERGM in general preserves the original network information better than the other two approaches with respect to important relationship summary statistics and statistical inferences from different network models. On top of the better utility, DP-ERGM also preserves the network privacy using the node DP, a stronger notion of privacy than the edge DP under which DWRR and SCEA operate.

Despite the better performance of DP-ERGM than DWRR and SCEA in general, there is room for improvement for DP-ERGM. For example, one may spend some privacy budget in choosing an ERGM using the data at hand rather than depending on external knowledge, compare different randomization mechanisms that draw posterior samples of the model parameters privately; or develop a hybrid private network generative model that would leverage the advantages DP-ERGM and SCAE. In addition, we are planning to look into coupling the DP-ERGM approach with the approaches for sanitizing nodal data and examine the utility of the fully synthesized networks and will continue to explore new approaches for synthesizing and releasing private networks. Finally, we would like to apply the synthesis techniques to other networks of various types and sizes.

References

[1] K. Granville, "Facebook and cambridge analytica: What you need to know as fallout widens," https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html, 2018.

[2] B. Mustanski, M. Birkett, L. M. Kuhns, C. A. Latkin, and S. Q. Muth, "The role of geographic and network factors in racial disparities in hiv among young men who have sex with men: an egocentric network study," *AIDS and Behavior*, vol. 19, no. 6, pp. 1037–1047, 2015.

[3] M. Birkett, L. M. Kuhns, C. Latkin, S. Muth, and B. Mustanski, "The sexual networks of racially diverse young men who have sex with men," *Archives of sexual behavior*, vol. 44, no. 7, pp. 1787–1797, 2015.

[4] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography*. Springer, 2006, pp. 265–284.

[5] J. Blocki, A. Blum, A. Datta, and O. Sheffet, "The johnson-lindenstrauss transform itself preserves differential privacy," in *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*. IEEE, 2012, pp. 410–419.

[6] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith, "Analyzing graphs with node differential privacy," in *Theory of Cryptography*. Springer, 2013, pp. 457–476.

[7] S. Chen and S. Zhou, "Recursive mechanism: towards node differential privacy and unrestricted joins," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 2013, pp. 653–664.

[8] S. Raskhodnikova and A. Smith, "Efficient lipschitz extensions for high-dimensional graph statistics and node private degree distributions," in *Proceedings of 57th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2016*. IEEE, 2016, pp. 495–504.

[9] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*. IEEE, 2007, pp. 94–103.

[10] V. Karwa and A. B. Slavkovic, "Inference using noisy degrees differentially private
β-model and synthetic graphs," *Annals of statistics*, vol. 44, no. 1, p. 87112, 2016.

[11] Y. Wang and X. Wu, "Preserving differential privacy in degree-correlation based graph generation," *Transactions on Data Privacy*, vol. 6, p. 127145, 2017.

[12] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, "Systematic topology analysis and generation using degree correlations," in *Proceedings of SIGCOMM 2006*, vol. 36, 2006, p. 13546.

[13] V. Karwa, P. N. Krivitsky, and A. B. Slavković, "Sharing social network data: differentially private estimation of exponential family random-graph models," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 66, no. 3, pp. 481–500, 2017.

[14] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[15] F. Liu, "Generalized gaussian mechanism for differential privacy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 747 – 756, 2019.

[16] C. Dwork, M. Naor, O. R. G. N. Rothblum, and S. Vadha, "On the complexity of differentially private data release: efficient algorithms and hardness results," in *STOC '09 Proceedings of the forty-first annual ACM symposium on Theory of computing Pages*, 2009, pp. 381–390.

[17] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," *IEEE ICDE 24th International Conference*, pp. 277 – 286, 2008.

[18] T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, "New specification for exponential random graph models," *Sociological Methodology*, vol. 36, pp. 99–153, 2006.

[19] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison, "Recent development in exponential random graph models for social networks," *Social Networks*, vol. 29, pp. 192–215, 2007.

[20] F. Liu, "Model-based differentially private data synthesis," *arXiv:1606.08052*, 2016.

[21] C. M. Bowen and F. Liu, "Comparative study of differentially private data synthesis methods," *Statistical Science*, 2020.

[22] H. Liu, I. H. Jin, Z. Zhang, and Y. Yuan, "Social network mediation analysis: a latent space approach," *arXiv*, vol. 1810.03751, 2018.

[23] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *Journal of the american Statistical association*, vol. 97, no. 460, pp. 1090–1098, 2002.

TABLE III: Latent space model parameter estimates based on synthetic friendship networks.

privacy budget	parameter (original)	θ_{GPA} (-0.169)			$\theta_{\text{\# cigarette}}$ (-0.084)			θ_{Gender} (1.197)			θ_{Class} (4.726)		
	DP-ERGM	**DWRR**	SCEA	**DP-ERGM**	DWRR	SCEA	**DP-ERGM**	DWRR	SCEA	DP-ERGM	DWRR	**SCEA**
average deviation from the original												
e^{-1}	0.249	**0.167**	-0.498	**0.077**	0.084	-0.124	**-0.509**	-1.185	1.180	-3.940	-4.665	**3.705**
e	0.239	**0.157**	-0.399	**0.079**	0.081	-0.100	**-0.488**	-1.118	0.909	-3.927	-4.268	**2.649**
e^2	0.230	**0.123**	-0.288	0.080	**0.072**	-0.075	**-0.479**	-0.943	0.633	-3.939	-3.385	**1.938**
root mean squared deviation												
e^{-1}	0.282	**0.169**	1.507	**0.079**	0.084	1.374	**0.522**	1.185	1.371	3.942	4.665	**3.726**
e	0.268	**0.159**	1.303	**0.081**	**0.081**	1.181	**0.500**	1.118	1.336	3.930	4.269	**2.836**
e^2	0.255	**0.127**	1.805	0.083	**0.072**	1.760	**0.492**	0.943	2.008	3.942	3.385	**1.969**

ϵ method degree ESP geodesic distance

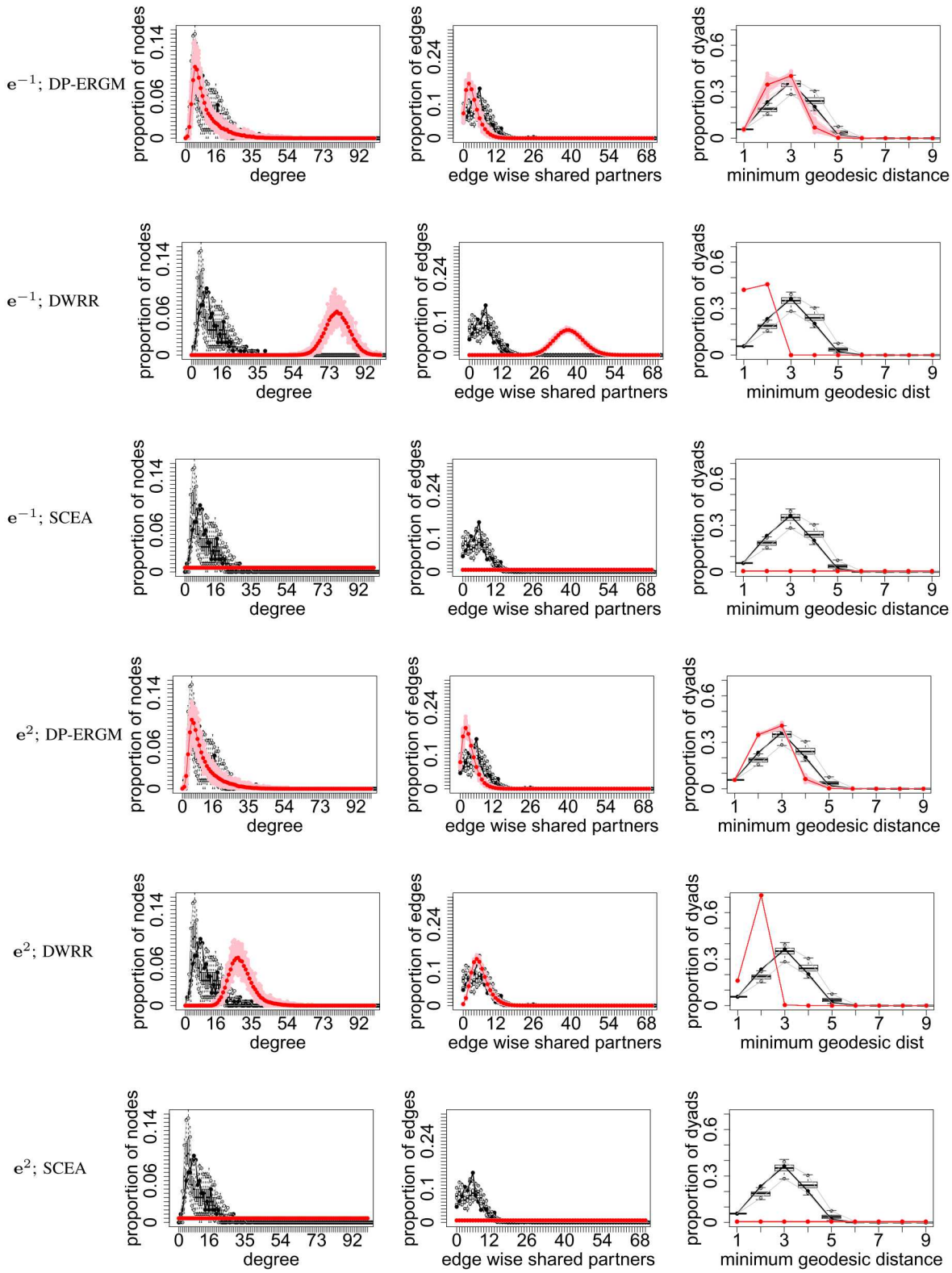

*Fig. 4: ERGM Goodness of Fit based on the synthetic college friendship networks. From left to right: proportion of nodes vs. degree, proportion of edges vs ESP, proportion of dyads vs geodesic distance. The **solid black lines** represent the statistics from the observed network; the black box plots represent the distributions of the statistics over 100 simulated networks given the original parameter estimates. The 100 pink lines (from 100 repetitions) in each plot represent the averaged statistics over 4 synthetic networks per repetition. The red lines represent the averages over the 100 repetitions.*

ϵ method degree ESP geodesic distance

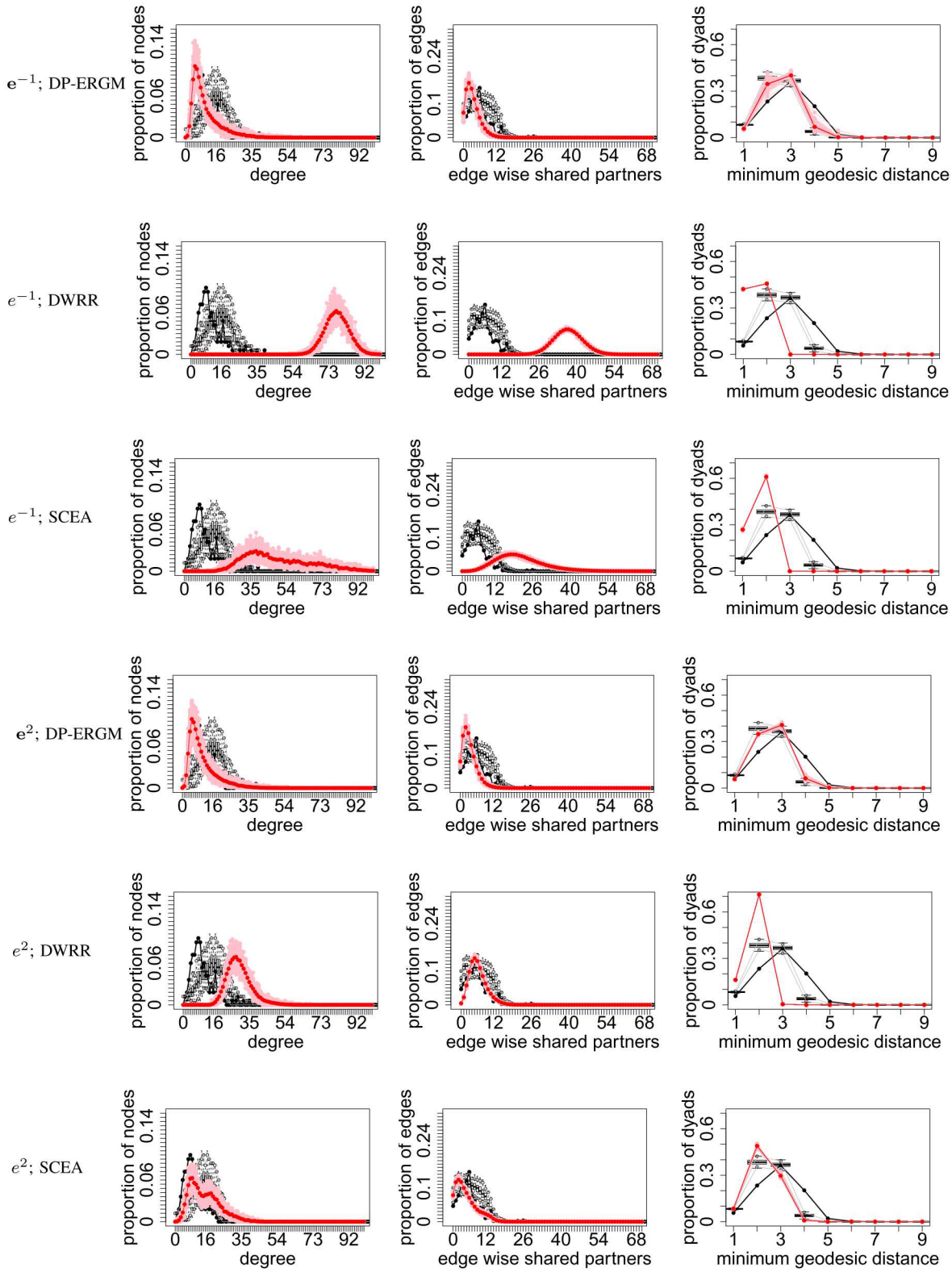

Fig. 5: Goodness of Fit for latent space model based on the synthetic college friendship networks (refer to Figure 4 caption for what each type of line represents in the plots).