Exponential Random Graph Estimation under Differential Privacy

Wentian Lu
Gerome Miklau

UMass Amherst
Privately analyzing social network

- Common graph statistics
  - Degree sequence [HLMJ09]
  - Triangle/star [KRSY11,NRS07]
  - Joint degree distribution [PGM12]
  - Clustering coefficient [WWZX13]

- Modeling graph
  - Kronecker model [MW12]
  - Exponential random graph?
Exponential random graph model (ERGM)

- The “state of art” in statistical social network modeling in the social sciences
- Support statistical inference on the processes influencing the formation of network structure
- Has been successfully applied in social science, e.g.
  - Co-sponsorship network [CD11]
  - Friendship network [GKM09]
  - Inter-organizational network, and more [LKR12]
Given: an observed network $G$, a set of ERGM model terms $f$

Output: estimated parameters for the model under differential privacy
Privacy definition: differential privacy

- **Differential privacy** [DMNS06]

  - $D$ and $D'$ are called *neighboring databases*
  - Provable privacy for protecting every *individual* while releasing aggregating properties of entire table

\[
\text{probability distribution on outputs is essentially unchanged}
\]
Background: differential privacy

- Formally,
  \[
  \Pr[K(D) \in O] \leq e^\varepsilon \Pr[K(D') \in O]
  \]
  - \(\varepsilon\): privacy parameter
  - smaller \(\varepsilon\), more privacy; bigger \(\varepsilon\), more utility.

- Output perturbation: Laplace mechanism
  - **Global Sensitivity** of a query \(f\)
    \[
    GS_f = \max_{\forall\text{ neighbors } D, D'} |f(D) - f(D')|
    \]
  - Adding noise calibrated to sensitivity
    \[
    K(D) = f(D) + Lap(0, GS_f / \varepsilon)
    \]
  - Larger GS \(\rightarrow\) more noise
ERGM: basics

- A parametric model defining probability
distributions over networks

\[ p(x | \theta) = \frac{\exp(\theta \cdot f(x))}{Z_\theta} \]

Let \( x_0 \) be the observed network. \( X \) be the random
variable representing network.

\[ \mathbb{E}_\theta(f(X)) = f(x_0) \]

Our goal: release \( \theta \) under differential privacy
ERGM: parameter estimation

- A maximum likelihood estimation problem

\[
\arg\max_\theta p(x_0|\theta)
\]

Example

\[f = \# \text{ edges in a network}\]

\[\theta = \log \left( \frac{\# \text{ edges in } x_0}{\binom{n}{2} - \# \text{ edges in } x_0} \right)\]

- However, most ERGMs don’t have analytical estimate

Very hard to compute sensitivity.
Contributions

ERGM specification

Sufficient statistics

Private sufficient statistics

Private estimated parameters

Chain mechanism

Estimation algorithm

No access to original network or true sufficient statistics.
ERGM: model terms \( f(x) \)

- **Traditional network statistics**
  - edges, stars, triangles, loops

- **Alternating statistics** [SPRH06]
  - k-star, k-triangle, k-twopath
  - More structural information and proved to lead to better modeling
Alternating statistics

3-star

2-triangle

2-twopath

**k-star:**

\[ S_k = \sum_i \binom{d_i}{k} \text{ where } d_i \text{ is the degree of node } i \]

**Alternating k-star:**

\[ S(x; \lambda) = S_2 - \frac{S_3}{\lambda} + \ldots + (-1)^{n-1} \frac{S_{n-1}}{\lambda^{n-3}} \]
Example

\[ M \sim \text{edges} + \text{alternate } k\text{-triangle} \]

1. edges
2. alternate \( k \) triangle

**Global sensitivity of alt \( k \)-triangle is \( O(n) \)!

1. (private) edges
2. (private) alternate \( k \)-triangle

\[ \theta = (\theta_1, \theta_2) \]
Differential privacy mechanism for addressing high global sensitivity

- **Local sensitivity** \([NRS07]\)

\[
\text{LS}_f(x) = \max_{\forall x', \text{ s.t. neighbors } x, x'} |f(x) - f(x')|
\]

However, local sensitivity cannot be used directly.
Chain mechanism

1. Add **Exponential** noise to $LS_f(x)$ with $\frac{\varepsilon}{2}$ using **global sensitivity of** $LS_f(x)$
2. Add **Laplace** noise $f(x)$ with $\frac{\varepsilon}{2}$ using bounded $LS_f(x)$ from step 1

Bounding local sensitivity:
1. The bound is not smaller than local sensitivity.
2. The bound itself is private.
Chain mechanism on real graphs with $\varepsilon = 0.1$

alt k-star

alt k-triangle

alt k-twopath

Relative RMSE

CA-HepTh
n=9.9k
m=26k

CA-HepPh
n=12k
m=119k

CA-AstroPh
n=19k
m=198k

CA-CondMat
n=23k
m=93k

Email-Enron
n=37k
m=184k
ERGM parameter estimation with private statistics

Standard estimation
- Private statistics as input
- Run standard estimation algorithm

In fact, we can do better than that!

Key observation:
Adapting the estimation process by incorporating the distribution of random noise.
Bayesian inference

Private statistics: $\tilde{y}$
The unknown network: $x$
ERGM parameter: $\theta$

Posterior: $p(\theta | \tilde{y})$

$$p(\theta | \tilde{y}) \propto p(\tilde{y} | \theta)p(\theta)$$

$$= \sum_x p(\tilde{y} | x)p(x | \theta)p(\theta)$$

Privacy distribution

ERGM definition

Prior distribution of parameter

Laplace mechanism
Chain mechanism

$$p(x | \theta) = \frac{\exp(\theta \cdot f(x))}{Z_\theta}$$
ERGM specifications:
M1 $\sim$ edges + alternating k-star
M2 $\sim$ edges + alternating k-triangle
M3 $\sim$ edges + alternating k-twopath

standard estimation $\theta = (\theta_1, \theta_2)$

Bayesian estimation $\theta = (\theta_1, \theta_2)$
Summary of contributions

- Solve the problem of estimating parameters for ERGM under differential privacy

- Approach: decomposing into two steps, releasing private statistics first and running estimation second.

- Our local sensitivity-based chain mechanism can offer lower error than existing methods.

- Bayesian inference based parameter estimation is flexible and more accurate than standard methods.