Private Analysis of Graph Structure

Grigory Yaroslavtsev
http://grigory.us

With Vishesh Karwa, Sofya Raskhodnikova and Adam Smith
Pennsylvania State University
Publishing network data

Many data sets can be represented as a graph:

- Friendship in online social network
- Financial transactions
- Romantic relationships

- Publish information about a graph
- Preserve privacy of relationships

Naïve approach: anonymization

American J. Sociology, Bearman, Moody, Stovel
Goal: Publish structural information about a graph

- Anonymization not sufficient [Backström, Dwork, Kleinberg ’07, Narayanan, Shmatikov ’09, Narayanan, Shi, Rubinstein ’11]
- Ideal: Algorithms with rigorous privacy guarantee, no assumptions about attacker’s prior information/algorithm
Differential privacy
[Dwork, McSherry, Nissim, Smith ’06]

- Limits **incremental** information by hiding presence/absence of an individual relationship

**Neighbors**: Graphs $G$ and $G'$ that differ in one edge

- Answers on neighboring graphs should be similar
**Differential privacy for relationships**

**$\epsilon$-differential privacy (edge privacy)**

For all pairs of neighbors $G, G'$ and all events $S$:

$$\Pr[A(G) \in S] \leq e^{\epsilon} \Pr[A(G') \in S]$$

- Probability is over the randomness of $A$
- Definition requires that the distributions are close:
Subgraph counts

For graphs $G$ and $H$: # of occurrences of $H$ in $G$

Example:

- **2-star:** Total: 40
- **Triangle:** Total: 2
- **2-triangle:** Total: 1

$\cdots$
Subgraph counts

- Subgraph counts are used in:
  - Exponential random graph models
  - Descriptive graph statistics, e.g.:

\[
\text{Clustering coefficient} = \frac{\# \begin{array}{c}
\text{triangle}
\end{array}}{\# \begin{array}{c}
\text{triplet}
\end{array}}
\]

- Our focus: efficient differentially private algorithms for releasing subgraph counts
Previous work

• **Smooth Sensitivity** [Nissim, Raskhodnikova, Smith ‘07]
  – Differentially private algorithm for triangles
  – **Open:** private algorithms for other subgraphs?

• Private queries with joins [Rastogi, Hay, Miklau, Suciu ‘09]
  – Works for a wide range of subgraphs
  – Weaker privacy guarantee, applies only for specific class of adversaries

• **Private degree sequence** [Hay, Li, Miklau, Jensen ’09]
  – Guarantees differential privacy
  – Works for k-stars, but not for other subgraphs
Laplace Mechanism and Sensitivity
[Dwork, McSherry, Nissim, Smith ‘06]

- Add noise: mean = 0, standard deviation \( \sim (S_f/\epsilon) \),
  where \( S_f \) is sensitivity => \( \epsilon \)-differential privacy:

  \[
  f'(G) = f(G) + \text{Lap}(S_f/\epsilon)
  \]

- Local sensitivity ([NRS’07], not differentially private!):

  \[
  LS_f(G) = \max_{G': \text{Neighbor of } G} |f(G) - f(G')|
  \]

- Previous work (mostly): Global sensitivity

  \[
  S_f = GS_f = \max_G LS_f(G) \Rightarrow \text{differentially private!}
  \]
Instance-Specific Noise

\( G_n \) = set of all graphs on \( n \) vertices. \( d(G,G') = \# \text{ edges in which } G \text{ and } G' \text{ differ.} \)

**Smooth Sensitivity** [Nissim, Raskhodnikova, Smith ’07]:

\[
S^*_f,\beta(G) = \max_{G' \in G_n} (LS_f(G') \cdot e^{-\beta d(G,G')})
\]

- Add Cauchy noise: median = 0, median absolute value \( \propto S^*_f,\beta(G)/\beta \) (where \( \beta = c \cdot \epsilon \)) \( \Rightarrow \epsilon \)-differential privacy:

\[
f'(G) = f(G) + \text{Cauchy}(S^*_f,\beta/\beta)
\]

- Naïve computation requires exponential time
- [NRS’07]: Compute smooth sensitivity for triangles
Our contributions

• Differentially private algorithms for k-stars and k-triangles
  – Efficiently compute smooth sensitivity for k-stars
  – \textbf{NP-hardness} for k-triangles and k-cycles
  – Different approach for k-triangles
• Average-case analysis in $G(n,p)$
• Theoretical comparison with previous work
• Experimental evaluation
Smooth Sensitivity for k-stars (\begin{figure}...

This paper: near-linear time algorithm for smooth sensitivity

• Algorithm also reveals structural results, e.g.:
  – **Proposition:**
    \[
    \text{If } (\epsilon < 1) \text{ and } (\text{maximum degree} > \text{const} \cdot k/\epsilon) \text{ then } \text{(smooth sensitivity)} = \text{(local sensitivity)}
    \]

• Algorithm optimal for large class of graphs
  – **Proposition:** error > const \cdot (local sensitivity)

• Compared to [HLMJ’09] (private degree sequence):
  – Our error never worse by more than a constant factor
  – For 2-stars, our error can be better by \Omega\left(\sqrt{n/\epsilon}\right) factor
Private Approximation to Local Sensitivity: k-triangles

Approximate differential privacy, \((\epsilon, \delta)\)-privacy

[Dwork, Kenthapadi, McSherry, Mironov, Naor ‘06]:

\[
\Pr[A(G) \in S] \leq e^\epsilon \Pr[A(G') \in S] + \delta
\]

Idea: Private upper bound on local sensitivity \((\widehat{LS})\).

Release: \(A(G) = (\widehat{LS}, f(G) + \text{Lap}(\widehat{LS}/\epsilon))\).

If

- \(\widehat{LS}\) is \(\epsilon\)-differentially private and
- \(\Pr[\widehat{LS} \geq LS] \geq 1 - \delta\)

Then \(A\) is \((2\epsilon, e^\epsilon \delta)\)-differentially private.
Evaluating our algorithms

• Theoretical evaluation in $G(n,p)$ model
  – All of our algorithms have relative error $\rightarrow 0$ when the average degree $= np$ grows

• Empirical evaluation
  – Synthetic graphs from $G(n,p)$ model
  – Variety of real data sets
Experimental results for $G(n,p)$

- Comparison with previous work for $p = \frac{\log n}{n}$

![Graph showing comparison of algorithms]

- 2-Stars
- LS Barrier
- Our algorithms
- HLMJ
- RHMS Lower
- RHMS upper
- Relative Error = 1
- 5% Relative Error
Experimental results for $G(n,p)$

- Comparison with previous work for $p = \frac{\log n}{n}$
Experimental results for $G(n,p)$

- Comparison with [RHMS’09] for $p = \frac{\log n}{n}$

![Graphs showing relative median error for Triangles and 2 Triangles with different comparison results.](image-url)
Experimental results (SNAP)

Relative Median Error

- 2-triangles
- triangles

<table>
<thead>
<tr>
<th>Dataset</th>
<th>n</th>
<th>m</th>
<th>Dataset</th>
<th>n</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>ca-GrQc</td>
<td>5K</td>
<td>29K</td>
<td>ca-HepTh</td>
<td>10K</td>
<td>52K</td>
</tr>
<tr>
<td>ca-CondMat</td>
<td>23K</td>
<td>187K</td>
<td>ca-HepPh</td>
<td>12K</td>
<td>237K</td>
</tr>
<tr>
<td>Email-Enron</td>
<td>37K</td>
<td>368K</td>
<td>ca-AstroPh</td>
<td>19K</td>
<td>396K</td>
</tr>
</tbody>
</table>
Experimental results (SNAP)

- 2-stars
- HLMJ - 2-stars

<table>
<thead>
<tr>
<th>Network</th>
<th>n</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>ca-GrQc</td>
<td>5K</td>
<td>29K</td>
</tr>
<tr>
<td>ca-HepTh</td>
<td>10K</td>
<td>52K</td>
</tr>
<tr>
<td>ca-CondMat</td>
<td>23K</td>
<td>187K</td>
</tr>
<tr>
<td>ca-HepPh</td>
<td>12K</td>
<td>237K</td>
</tr>
<tr>
<td>Email-Enron</td>
<td>37K</td>
<td>368K</td>
</tr>
<tr>
<td>ca-AstroPh</td>
<td>19K</td>
<td>396K</td>
</tr>
</tbody>
</table>

Relative Median Error

2-stars vs HLMJ - 2-stars comparison for various networks with different numbers of nodes (n) and edges (m).
Summary

• Private algorithms for subgraph counts
  – Rigorous privacy guarantee (differential privacy)
  – Running time close to best algorithms for computing the subgraph counts

• Improvement in accuracy and (for some graph counts) in privacy over previous work

• Techniques:
  – Fast computation of smooth sensitivity
  – Differentially private upper bound on local sensitivity