Zero-Knowledge Private Graph Summarization

Maryam Shoaran
Alex Thomo
Jens Weber

University of Victoria, Canada
Outline

- Introduction
- Challenge: Evidence of Participation
- Sample Aggregates
- Zero-Knowledge Privacy
- Analysis of Utility of ZKP
- Conclusions
Privacy of Aggregate Information

- Aggregate query \( q : D \rightarrow R \)

- **Background knowledge** can help infer sensitive information about participants from aggregate query answers.
Example

- Healthcare data in a hospital:
  - Aggregate query
    - What is the number of patients with cancer diagnosis admitted today?
    - Answer=2.
  - Background knowledge:
    - Alice was admitted today.
    - 6 patients in total were admitted today.

Alice has cancer with probability 1/3.
Randomize the algorithm, so that it has a probability distribution over outputs such that if a person removed his/her input, the relative probabilities of any output don’t change by much.

Can pretend your input does not data about a given person.

Can view as model of “plausible deniability”.
Definition: Randomized algorithm $San$ satisfies $\epsilon$-DP

equivalent to

for any two neighboring databases $D$ and $D'$

$$Pr[ San(D) \in W ] \leq e^\epsilon \times Pr[ San(D') \in W ]$$
Typical way to achieve DP:
- Add **properly calibrated** Laplace noise to query answer.
  - Sanitized output: $San(D) = q(D) + \text{noise}$,
  - PDF of Laplace Noise with mean zero:
    \[
    h(x) = \frac{1}{2\lambda} e^{-\frac{|x|}{\lambda}}
    \]

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith (TCC 2006)
Differential Privacy (III)

- Sensitivity of $q : D \rightarrow R$

$$\Delta(q) = \max_{D,D'} | q(D) - q(D') |$$

- Calibrate noise scale $\lambda$ to the sensitivity of the query:

$$\lambda = \frac{\Delta(q)}{\varepsilon}$$
Problem of DP for Social Networks
Problem of DP for Social Networks
Problem of DP for Social Networks

We can still guess that Bob is friend with Alice!

DP doesn't protect against evidence of participation.
Problem of DP for Social Networks

- DP ensures that for any true answer, $c$ or $c - 1$, the sanitized answer is pretty much the same.

- However, not strong enough:
  - Existence of Bob’s edge changes the true answer not just by 1, but by a bigger number
    - as it causes more edges to be created
ZKP Intuition

- ZKP guarantees that an attacker cannot discover
  - any personal information
  - more than
  - what can be inferred from some aggregate on a sample of a database with the person removed.

Suppose the network size is $10,000$ and the sample size is $\sqrt{10,000} = 100$.

- Evidence provided by the 7 more edges caused by Bob’s edge will essentially be protected;
- With a high probability, none of these 7 edges will be in the sample.
Sample Complexity of a Function

\( \Pr(\left| T(D) - q(D) \right| \leq \delta) \geq 1 - \beta \)

- \((\delta, \beta)\)-sample complexity (SC) of \( q \).
- \( \delta \) is the sample error
Recall Sensitivity of a Function

- Sensitivity of $q : D \rightarrow R$
  \[ \Delta(q) = \max_{D,D'} |q(D) - q(D')| \]

- In DP we calibrate Laplace noise scale $\lambda$ to the sensitivity of the query:
  \[ \lambda = \frac{\Delta(q)}{\varepsilon} \]

- In ZKP we again use Laplace noise, but also consider the sample complexity of $q$.
  \[ \lambda = \frac{\Delta(q) + \delta}{\varepsilon} \]
Definition: A randomized algorithm $San$ satisfies $\epsilon$-ZKP w.r.t. sample aggregate $T$

iff

for any two neighboring databases $D$ and $D'$

$$\Pr[\text{Adv}(San(D), z) \in W] \leq e^\epsilon \times \Pr[\text{Sim}(T(D'), z) \in W]$$

$$\Pr[\text{Sim}(T(D'), z) \in W] \leq e^\epsilon \times \Pr[\text{Adv}(San(D), z) \in W]$$
Theorem [GLP11]

$q: G \rightarrow [a,b]^m$ has $(\delta, \beta)$-sample complexity w.r.t. $T$.

Then, $San(G) = q(G) + (X_1, ..., X_m)$ is

$$\ln \left( (1 - \beta) e^{\frac{\Delta(q) + \delta}{\lambda}} + \beta e^{\frac{(b-a)m}{\lambda}} \right) - ZKP$$

w.r.t. $T$. 

12/17/2015  
University of Victoria  
18
Graph Summarization

\[ g' \]

\[ g'' \]

\[ w_1 \]

\[ w_2[x] \quad w_2[y] \quad w_2[z] \]

\[ (.75, .33, 1) \]

\[ w_1 \]

\[ .4 \]

\[ .6 \]
## Results

<table>
<thead>
<tr>
<th>Term</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta(w_1)$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\Delta(w_2[x])$</td>
<td>$\frac{1}{r}$</td>
</tr>
<tr>
<td>$\Delta(w_2[y])$</td>
<td>$\frac{1}{r^2}$</td>
</tr>
<tr>
<td>$\Delta(w_2[z])$</td>
<td>$\frac{1}{r}$</td>
</tr>
</tbody>
</table>

- $w_1: \left( \delta, 2e^{-2k \delta^2} \right) - SC$
- $w_2[x]: \left( \delta, 2e^{-2k_g \delta^2} \right) - SC$
- $w_2[z]: \left( \delta, 2e^{-2k_g \delta^2} \right) - SC$
- $w_2[y]: \left( \delta, 2e^{-2k_g \times k_g \delta^2} \right) - SC$

- Smallest allowed group size
- $k$ is the sample size
- $k_g$ is the size of $g$ in a sample of size $k$
Results

Considering

\[ k = \sqrt[3]{n^2} \]
\[ \delta = \frac{1}{\sqrt[3]{n^2}} \]
\[ \lambda = \frac{\Delta(q) + \delta}{\varepsilon} \]

and using the ZKP theorem we get for \( w_1 \):

By adding noise

\[ \text{Lap}\left( \frac{1}{\varepsilon \cdot \sqrt[3]{k}} \right) \]

we have a San that is:

\[ \ln\left( \varepsilon + 2e^{-\sqrt[3]{k}} \right) - \text{ZKP} \]
Results

Considering

\[ k = \sqrt[3]{n^2} \]

\[ \delta = \frac{1}{\sqrt[3]{n^2}} \]

\[ \lambda = \frac{\Delta(q) + \delta}{\epsilon} \]

and using the ZKP theorem we get for \( w_2[x] \):

By adding noise

\[ \text{Lap}\left( \frac{1}{\epsilon \cdot r} + \frac{1}{\epsilon \cdot \sqrt[3]{k_g}} \right) \]

we have a San that is:

\[ \ln\left( \epsilon + 2e^{-\sqrt[3]{k_g}} \right) \text{- ZKP} \]
Relationship between noise scale and database size

For $\lambda=0.1$, the probability that noise is between -0.15 and 0.15 is about 80%.

For $\lambda=0.15$, the probability that noise is between -0.15 and 0.15 is about 63%.

For $\lambda=0.2$, the probability that noise is between -0.15 and 0.15 is about 52%.

For: $\varepsilon=0.1$, $\delta = \frac{1}{\sqrt[3]{k}}$
Conclusions

- Showed how to use ZKP for graph summarization
- Showed when it is reasonable to use ZKP

**Upshot:**
- ZKP is quite useful for protecting not only the participation of a connection, but also the evidence of its participation.
- However, from a utility point of view, ZKP can only be applied meaningfully on big social graphs.
Thank you!
References

- Nasrin Hassanlou, Maryam Shoaran, Alex Thomo. Probabilistic Graph Summarization. WAIM 2013: 545-556