Enabling Accurate Analysis of Private Network Data

Michael Hay

Joint work with

Gerome Miklau, David Jensen, Chao Li, Don Towsley
University of Massachusetts, Amherst

Vibhor Rastogi, Dan Suciu
University of Washington

October 8, 2009
Analysis of social networks

Can we enable analysts to study networks in a way that protects sensitive information about participants?

Social network derived from mobile phone call records [Onnela, PNAS 07]
How to achieve both privacy and utility?

DATA OWNER

ANALYST

Q what is diameter?

Q what is maximum degree?

Q how many 3 cliques?

Q is Alice connected to Bob?

Private network

Allow aggregate statistics
provided facts about individuals are not disclosed
Query answer perturbation

$$A = Q(G) + \text{noise}$$

[Dwork, TCC 06]
Query answer perturbation

DATA OWNER

G

Alice  Bob  Carol
Dave  Ed
Fred  Greg  Harry

ANALYST

Q  how many edges?
A = Q(G) + noise

[Dwork, TCC 06]
Query answer perturbation

\[ \text{DATA OWNER} \]

\[ G \]

Alice  Bob  Carol
Dave  Ed
Fred  Greg  Harry

\[ \text{ANALYST} \]

Q  *how many edges?*

A = Q(G) + noise

11

[Dwork, TCC 06]
Query answer perturbation

\[ A = Q(G) + \text{noise} \]

\[ 11 + 2.3 \]

[Dwork, TCC 06]
Query answer perturbation

**DATA OWNER**

Alice  Bob  Carol

Dave  Ed

Fred  Greg  Harry

**ANALYST**

Q  how many edges?

A = Q(G) + noise

11 + 2.3 → 13.3

[Dwork, TCC 06]
Query answer perturbation

\[ \text{DATA OWNER} \]

\[ \text{ANALYST} \]

\[ G \]

Alice \quad Bob \quad Carol

Dave \quad Ed

Fred \quad Greg \quad Harry

\[ Q \text{ how many edges?} \]

\[ A = Q(G) + \text{noise} \]

11 + 2.3 \rightarrow 13.3

\[ \text{[Dwork, TCC 06]} \]

- Dwork, McSherry, Nissim, Smith [Dwork, TCC 06] have described an answer perturbation mechanism satisfying differential privacy.

- Comparatively few results for these techniques applied to graphs.
Query answer perturbation

\[ G \]

DATA OWNER

Alice  Bob  Carol  Dave  Ed  Fred  Greg  Harry

ANALYST

Q: how many edges?

A = Q(G) + noise

11 + 2.3 \rightarrow 13.3

[Dwork, TCC 06]
Query answer perturbation

A = Q(G) + noise

\[[Dwork, TCC 06]\]
Query answer perturbation

\[ A = Q(G) + \text{noise} \]

\[ A = Q(G') + \text{noise} \]

\[ \text{DATA OWNER} \]

\[ \text{ANALYST} \]

\[ \text{Q} \quad \text{how many edges?} \]

\[ \text{Q} \]

\[ \text{A} = Q(G) + \text{noise} \]

\[ \text{A} = Q(G') + \text{noise} \]
Query answer perturbation

**DATA OWNER**

\[ A = Q(G) + \text{noise} \]

**ANALYST**

\[ \text{Pr}[ A = x \mid \mu = Q(G) ] \]

\[ Q \text{ how many edges?} \]

\[ A = Q(G') + \text{noise} \]
Query answer perturbation

\[ A = \mu = Q(G) \]

**DATA OWNER**

Alice  | Bob  | Carol
---|---|---
Dave  | Ed  |
Fred  | Greg | Harry

**ANALYST**

\[ Q \text{ how many edges?} \]

\[ A = Q(G) + \text{noise} \]

\[ \Pr[ A = x \mid \mu = Q(G) ] = p \]

\[ Q \]

\[ A = Q(G') + \text{noise} \]

[Dwork, TCC 06]
Query answer perturbation

DATA OWNER

G

Alice  Bob  Carol
Dave  Ed
Fred  Greg  Harry

G’

Alice  Bob  Carol
Dave  Ed
Fred  Greg  Harry

ANALYST

Q  how many edges?

A = Q(G) + noise

Pr[ A = x | \mu = Q(G) ] = p

Q

A = Q(G’) + noise

Pr[ A = x | \mu = Q(G’) ]
Query answer perturbation

\[ G \]

\[ G' \]

**DATA OWNER**

Alice  Bob  Carol

Dave  Ed

Fred  Greg  Harry

**ANALYST**

\[ A = Q(G) + \text{noise} \]

Pr\[ A = x | \mu = Q(G) \] = p

\[ A = Q(G') + \text{noise} \]

Pr\[ A = x | \mu = Q(G') \] = q

**how many edges?**

[Dwork, TCC 06]
Query answer perturbation

\[ A = Q(G) + \text{noise} \]

Pr[ \( A = x \mid \mu = Q(G) \) ] = p

Pr[ \( A = x \mid \mu = Q(G') \) ] = q

G

Alice Bob Carol
Dave Ed
Fred Greg Harry

G'

Alice Bob Carol
Dave Ed
Fred Greg Harry

Q \text{ how many edges?}

\[ A = Q(G') + \text{noise} \]

\[ \text{differ by at most factor of } e^\varepsilon \]

[Dwork, TCC 06]
Calibrating noise

The following algorithm for answering $Q$ is $\varepsilon$-differentially private:

$$A \xrightarrow{Q} Q(G) + \text{Noise}(\Delta Q / \varepsilon)$$

$\Delta Q$: Max change in $Q$, over any two graphs differing by single edge
Calibrating noise

• The following algorithm for answering $Q$ is $\varepsilon$-differentially private:

$$A \rightarrow Q \rightarrow Q(G) + \text{Noise}(\Delta Q / \varepsilon)$$

$\Delta Q$: Max change in $Q$, over any two graphs differing by single edge
Calibrating noise

- The following algorithm for answering $Q$ is $\epsilon$-differentially private:

$$Q(G) + \text{Noise}(\Delta Q / \epsilon)$$

$\Delta Q$: Max change in $Q$, over any two graphs differing by single edge
Calibrating noise

• The following algorithm for answering $Q$ is $\varepsilon$-differentially private:

$$Q(G) + \text{Noise}(\Delta Q / \varepsilon)$$

$\Delta Q$: Max change in $Q$, over any two graphs differing by single edge
Calibrating noise

- The following algorithm for answering Q is $\varepsilon$-differentially private:

$$Q(G) + \text{Noise}(\frac{\Delta Q}{\varepsilon})$$

$\Delta Q$: Max change in Q, over any two graphs differing by single edge

$\text{sensitivity of Q}$

$\text{true answer}$

$\text{privacy parameter}$

$\text{sample from Laplace distribution}$
Positive results in differential privacy

• Some common analyses have low sensitivity: contingency tables, histograms [Dwork, TCC 06]

• Data mining algorithms implemented using only low sensitivity queries: PCA, k-Means, Decision Trees [Blum, PODS 05]

• Learning theory: possible to learn any concept class with polynomial VC dimension; half-space queries can be learned efficiently [Blum, STOC 08]

• Many challenges remain...
  
  • Beyond tabular data

  • Optimal query strategies?
Accurate degree distribution estimation is possible

• Degree distribution: the frequency of each degree in graph.

• A widely studied property of networks.

\[ [1,1,2,2,4,4,4,4] \]
Accurate degree distribution estimation is possible

• Degree distribution: the frequency of each degree in graph.

• A widely studied property of networks.

![Graph G with nodes labeled Alice, Bob, Carol, Dave, Ed, Fred, Greg, Harry. Degree sequence: [1, 1, 2, 2, 4, 4, 4, 4, 4].]

Histogram

Degree sequence as a vector

[1, 1, 2, 2, 4, 4, 4, 4]

CCDF
Accurate degree distribution estimation is possible

- Degree distribution: the frequency of each degree in graph.

- A widely studied property of networks.

$$[1,1,2,2,4,4,4,4]$$

**G**

[Image of a graph with nodes labeled Alice, Bob, Carol, Dave, Ed, Fred, Greg, Harry]

[Histogram and CCDF graphs showing degree distribution]
Accurate degree distribution estimation is possible

- Degree distribution: the frequency of each degree in graph.

- A widely studied property of networks.

```
[1,1,2,2,4,4,4,4]
```

Degree sequence as a vector

Histogram

CCDF

Degree sequence
Accurate degree distribution estimation is possible

- Degree distribution: the frequency of each degree in graph.

- A widely studied property of networks.

![Diagram of a network with labeled nodes Alice, Bob, Carol, Dave, Ed, Fred, Greg, Harry. The degree sequence is shown as a vector: [1, 1, 2, 2, 4, 4, 4, 4]. A histogram and complementary cumulative distribution function (CCDF) are also depicted.]
Using the sort constraint

\[ S(G) = [10, 10, ....10, 10, 14, 18,18,18,18] \]
Using the sort constraint

\[ S(G) = [10, 10, \ldots, 10, 10, 14, 18, 18, 18, 18, 18] \]
Using the sort constraint

- S(G) true degree sequence
- noisy observations ($\epsilon = 2$)
- inferred degree sequence
• The output of the sorted degree query is not (in general) sorted.

• We derive a new sequence by computing the closest non-decreasing sequence: i.e. minimizing L2 distance.
The output of the sorted degree query is not (in general) sorted.

We derive a new sequence by computing the closest non-decreasing sequence: i.e. minimizing L2 distance.
• The output of the sorted degree query is not (in general) sorted.

• We derive a new sequence by computing the closest non-decreasing sequence: i.e. minimizing L2 distance.
• The output of the sorted degree query is not (in general) sorted.

• We derive a new sequence by computing the closest non-decreasing sequence: i.e. minimizing $L_2$ distance.
Experimental results

livejournal
n=5.3M

orkut
n=3.1M

powerlaw
α=1.5, n=5M

ε=.001

ε=.01
Standard Laplace noise is sufficient but not necessary for differential privacy.

By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.

Improvement in accuracy will depend on sequence.

After inference, noise only where needed.
After inference, noise only where needed

- Standard Laplace noise is sufficient *but not necessary* for differential privacy.
- By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.
  - Improvement in accuracy will depend on sequence
• Standard Laplace noise is sufficient but not necessary for differential privacy.

• By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.

  • Improvement in accuracy will depend on sequence
After inference, noise only where needed

- Standard Laplace noise is sufficient but not necessary for differential privacy.
- By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.
  - Improvement in accuracy will depend on sequence
Standard Laplace noise is sufficient but not necessary for differential privacy.

By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.

Improvement in accuracy will depend on sequence.
Conclusion

• Possible to accurately estimate degree distributions while providing strong guarantees of privacy

• Other findings

  • Some network analyses cannot be accurately answered under differential privacy (clustering coefficient, motif analysis [Nissim, STOC 07] [PODS 09])

  • Apply inference to other queries (e.g. histograms [CoRR 09])

• Future work: generate accurate synthetic networks under differential privacy?
Questions?

Additional details on our work may be found here:


http://www.cs.umass.edu/~mhay/
References


References (con’t)


