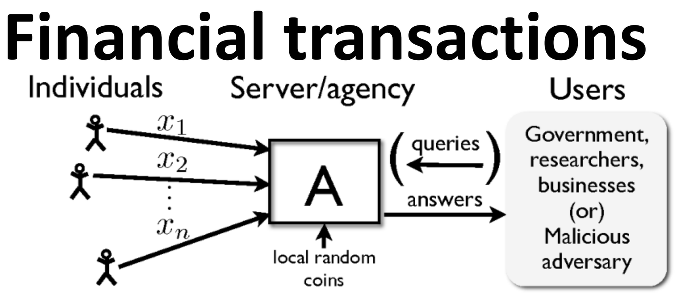
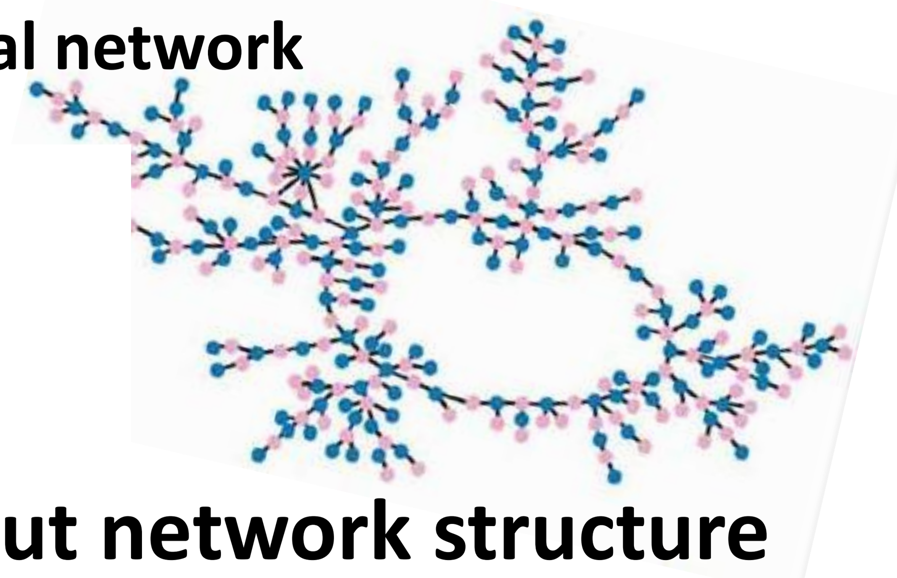
Private Analysis of Graph Stucture

V. Karwa, S. Raskhodnkova, A.Smith. G. Yaroslavtsev (VLDB'11)

Network Data

- **Many data sets can be naturally represented as network of pairwise relationships (graph)**
  - **Friendship in online social network**
  - **Romantic relationships**
  - **Financial transactions**





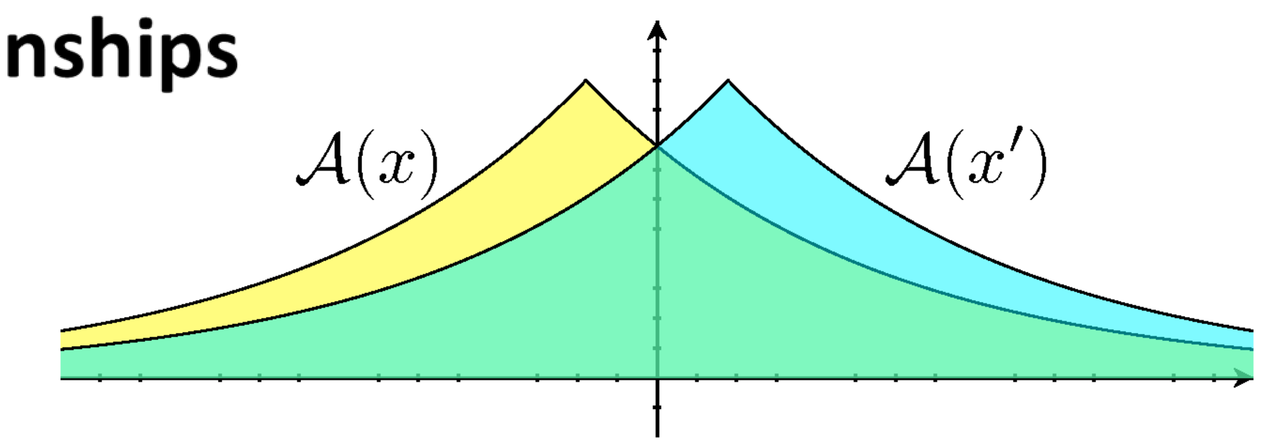
- **Goal: Publish info about network structure for use by researchers, gov't agencies, …**
- **Privacy concerns: Sensitive information**
  - **De-anonymizing social networks [BDK'07, NS'09]**
  - **Link prediction by de-anonymization [NSR'11]**
- **This work: new privacy-preserving algorithms**

Measuring Privacy

- **Goal: Rigorous privacy guarantee**
- **Challenges**
  - **Attacker's prior information unknown**
  - **Attack algorithm / methodology unknown**
- **ε-Differential privacy [DMNS'06]: For all neighbors G and G' (differing in one record):**

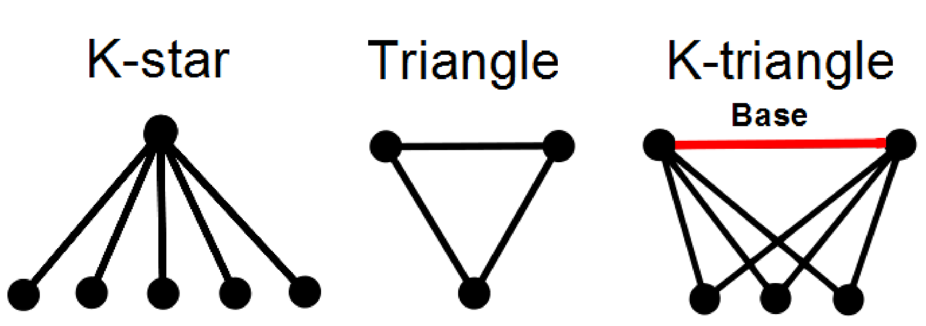
$$Pr[A(G) = s] \le e^{\varepsilon} \cdot Pr[A(G') = s]$$

- **Hides the presence/absence of individual relationships**



Subgraph Counts

- **For graphs G and H, $f_H(G)$ = # of occurrences of H in G**



- **Subgraph counts are widely used in:**
  - **Exponential random graph models**
  - **Descriptive graph statistics (e.g. clustering coefficient)**
- **Our work: We present d.p. algorithms for accurately releasing subgraph counts**

Techniques

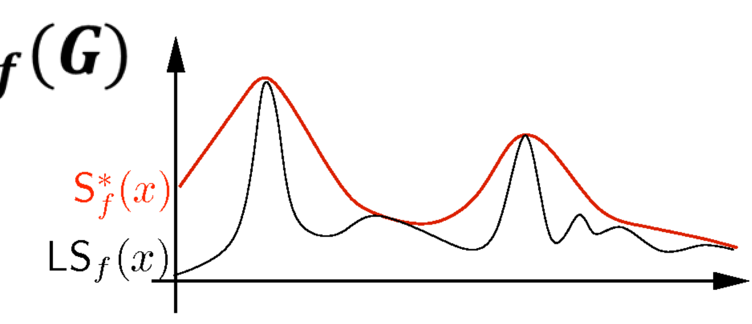
- **Add noise, proportional to S/ε (where S is sensitivity). How to minimize S?**
- **This work: Instance-specific noise**
  - **Local sensitivity (also used in [RHMS'09])**

$$LS_f(G) = \max_{G' \in N(G)} |f(G) - f(G')|$$

  - **Nearly optimal amplitude, but not d.p.**
- **Previous work: Global sensitivity (mostly)**

$$GS_f(G) = \max_G LS_f(G)$$

- **Main approaches:**
  - **Smooth sensitivity [NRS'07]**
  - **D.p. approximation to $LS_f(G)$ (this work).**



Results

- **New algorithms:**
  - **K-stars (linear time algorithms for computing smooth sensitivity)**
  - **K-triangles (d.p. approximation to $LS_f(G)$).**
- **Bounds on noise in terms of $LS_f(G)$:**
  - **Conditions, when amplitude of noise is close to $\frac{LS_f}{\varepsilon}$**
- **Average-case analysis in Gilbert model G(n,p) $(np = \omega(\sqrt{n} \log n), \varepsilon = o(\log n))$:**

$$\frac{|A_H(G) - f_H(G)|}{f_H(G)} = O\left(\frac{1}{\varepsilon n^2 p}\right)$$

- **Computing smooth sensitivity is NP-hard for some subgraph queries (K-triangles, K-cycles)**

Experimental Evaluation

Results of experimental evaluation for $\varepsilon = 0.5$:

Dataset	H	Median Relative Error
CaGrQc	△	0.081
$Nodes = 5242$	2⋆	0.0041
$Edges = 2890$	3⋆	0.144
CaHepTh	△	0.11
$Nodes = 9877$	2⋆	0.002
$Edges = 51971$	3⋆	0.0009
CaAstroPh	△	0.017
$Nodes = 18772$	2⋆	0.00043
$Edges = 396160$	3⋆	0.0023
CaHepPh	△	0.007
$Nodes = 12008$	2⋆	0.00038
$Edges = 237010$	3⋆	0.001
CaCondMat	△	0.077
$Nodes = 23133$	2⋆	0.00158
$Edges = 186936$	3⋆	0.011
Enron	△	0.032
$Nodes = 36692$	2⋆	0.00055
$Edges = 267662$	3⋆	0.0021

Supported by the National Science Foundation