Analyzing Private Network Data Using Output Perturbation

Gerome Miklau

Joint work with

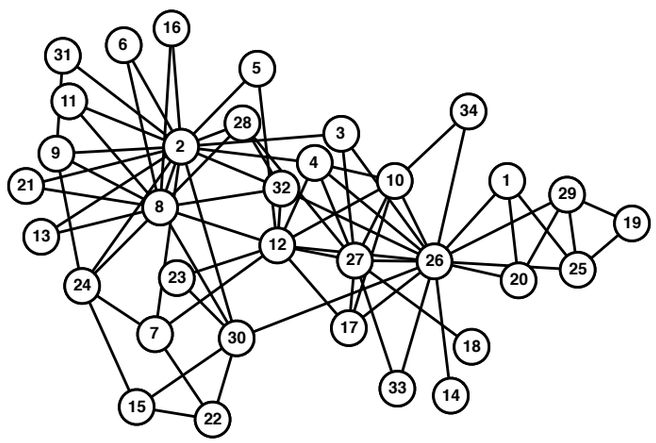
Michael Hay, David Jensen, Don Towsley
University of Massachusetts, Amherst

Vibhor Rastogi, Dan Suciu
University of Washington

September 11, 2009

Social network analysis -- transformed

1977



2008

“Zachary’s Karate Club”

W. W. Zachary
An information flow model for conflict and fission in small groups
Journal of Anthropological Research

34 nodes
78 edges

Global IM Network

J. Leskovec and E. Horvitz.
Planetary-scale views on a large instant-messaging network
Conference on the World Wide Web

180 million nodes
1.3 billion edges

The opportunity is not being realized

Major obstacle to data dissemination: privacy risk.

- Common outcomes:
  - No availability
  - Limited availability
    - only within institutions who own the data, or a among limited set of researchers who have negotiated access.
  - Availability, at a cost
    - Privacy of participants may be violated, bias or inaccuracy in released data.

Analysis of private networks

Can we permit analysts to study networks without revealing sensitive information about participants?

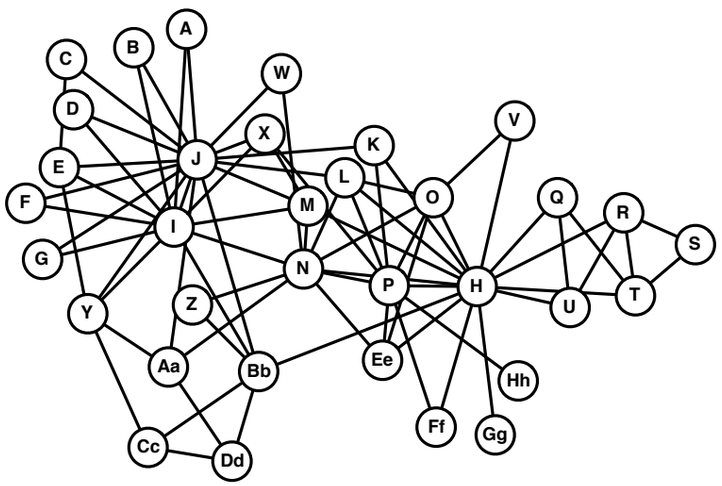
Example analyses based on network topology:

- **Properties of the degree distribution**
- **Motif analysis**
- Community structure
- Processes on networks: routing, rumors, infection
- Resiliency

Synthetic data release

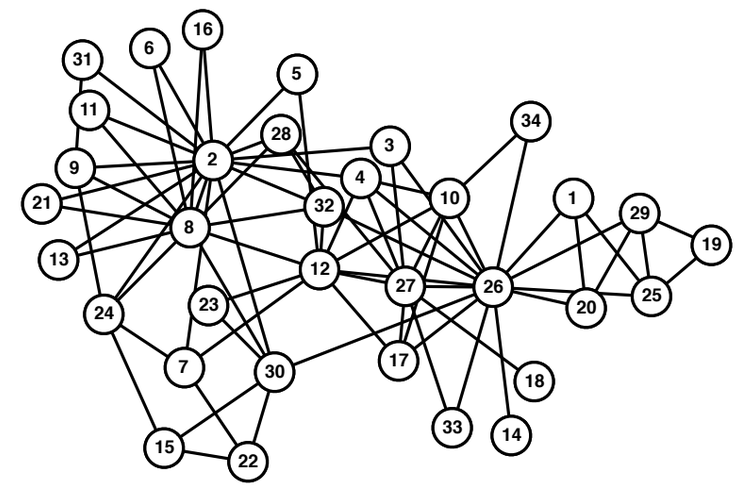
DATA OWNER

ANALYST



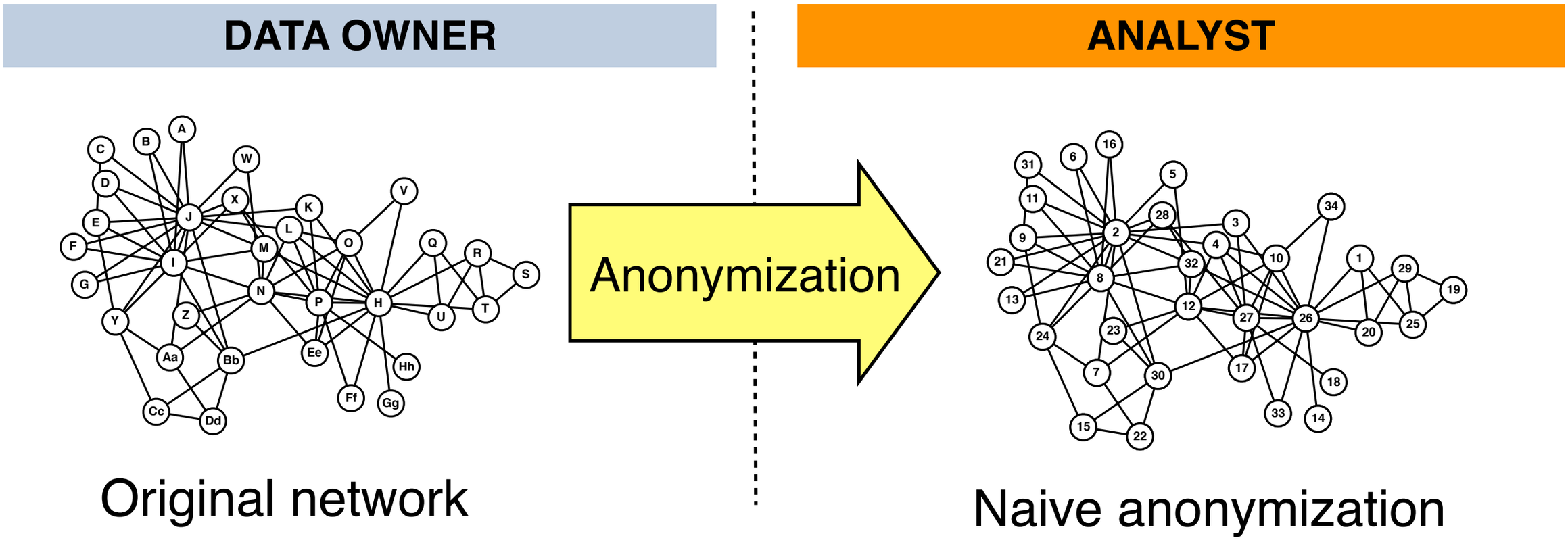
Original network

Anonymization



Naive anonymization

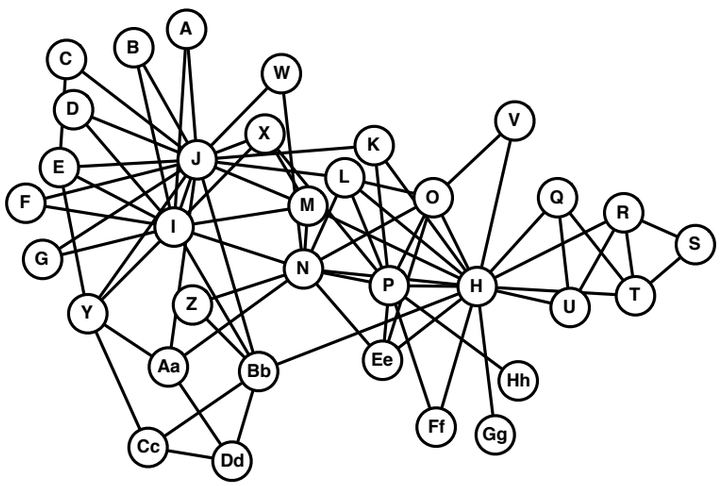
Synthetic data release



- Create topological similarity **[Liu, SIGMOD 08] [Zhou, ICDE 08] [Zou, VLDB 09]**

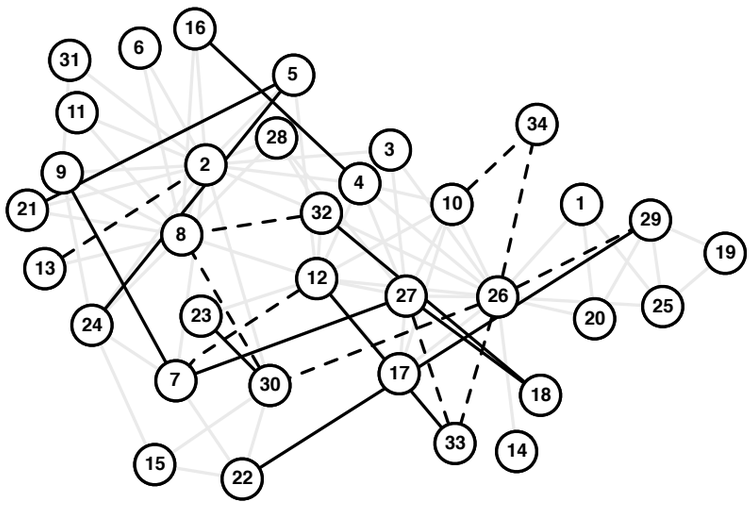
Synthetic data release

DATA OWNER | **ANALYST**



Original network

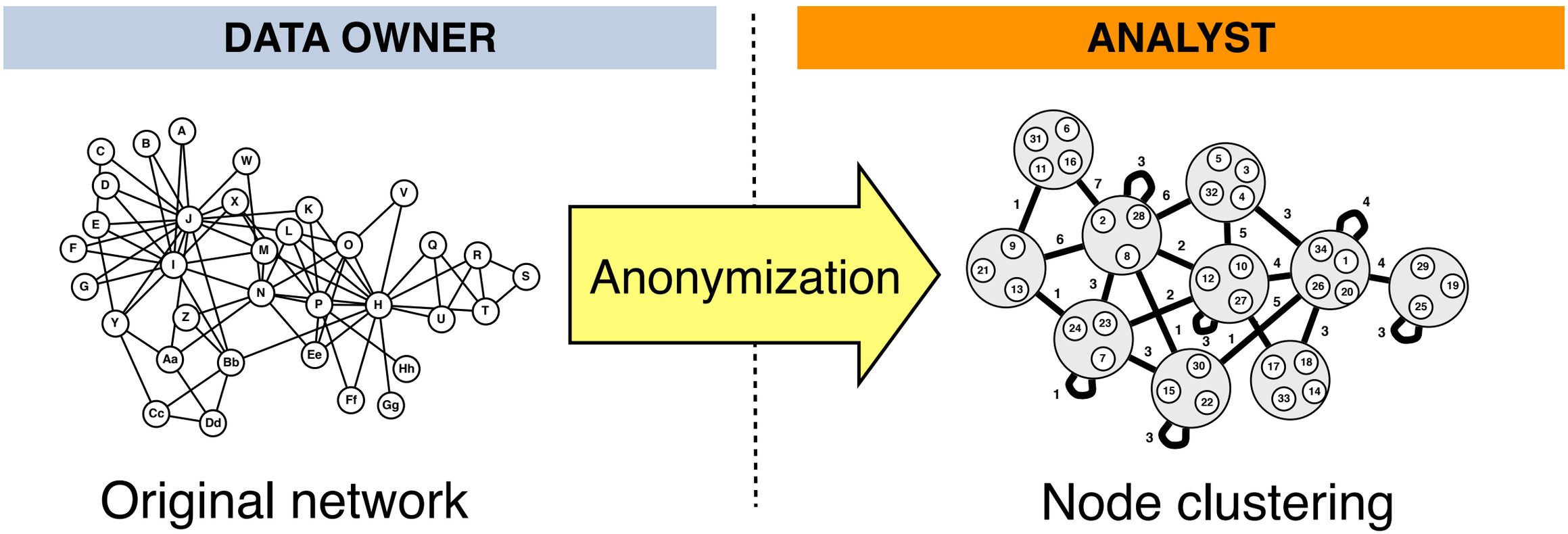
Anonymization



Randomized Edges

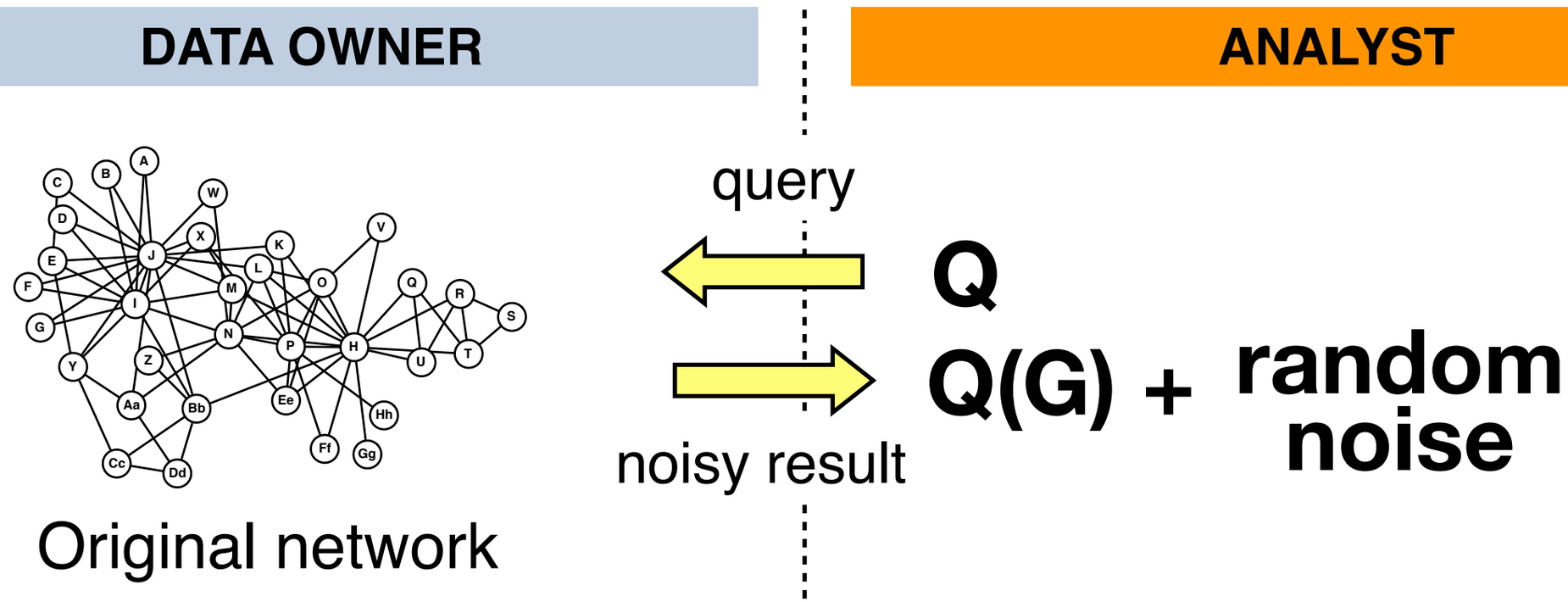
- Create topological similarity **[Liu, SIGMOD 08] [Zhou, ICDE 08] [Zou, VLDB 09]**
- Randomize edges **[Rastogi, VLDB 07] [Ying, SDM 2008]**

Synthetic data release



- Create topological similarity **[Liu, SIGMOD 08] [Zhou, ICDE 08] [Zou, VLDB 09]**
- Randomize edges **[Rastogi, VLDB 07] [Ying, SDM 2008]**
- Clustering/summarization **[Campan, PinKDD 08]** **[Hay, VLDB 08]** **[Cormode, VLDB 08] [Cormode, VLDB 09]**

Output perturbation

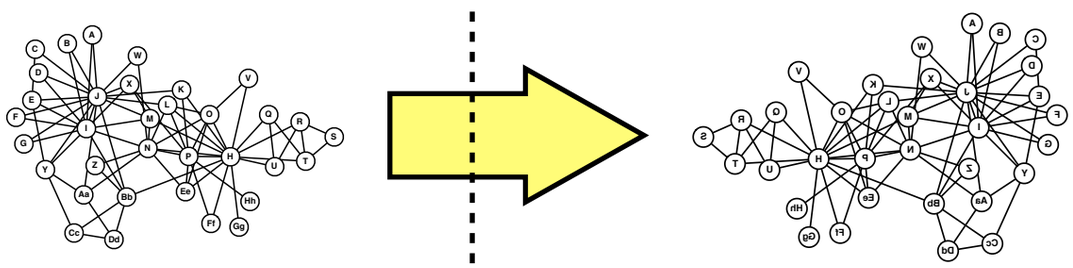


- Dwork, McSherry, Nissim, Smith **[Dwork, TCC 06]** have described an output perturbation mechanism satisfying ***differential privacy.***

- Comparatively few results for these techniques applied to graphs.

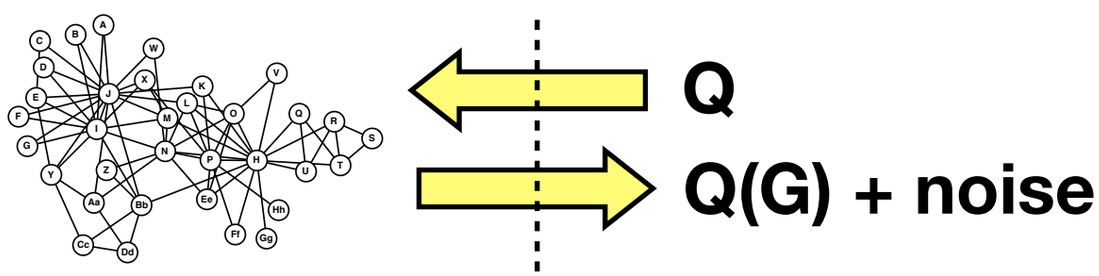
Synthetic data release v. output perturbation

- Synthetic data release



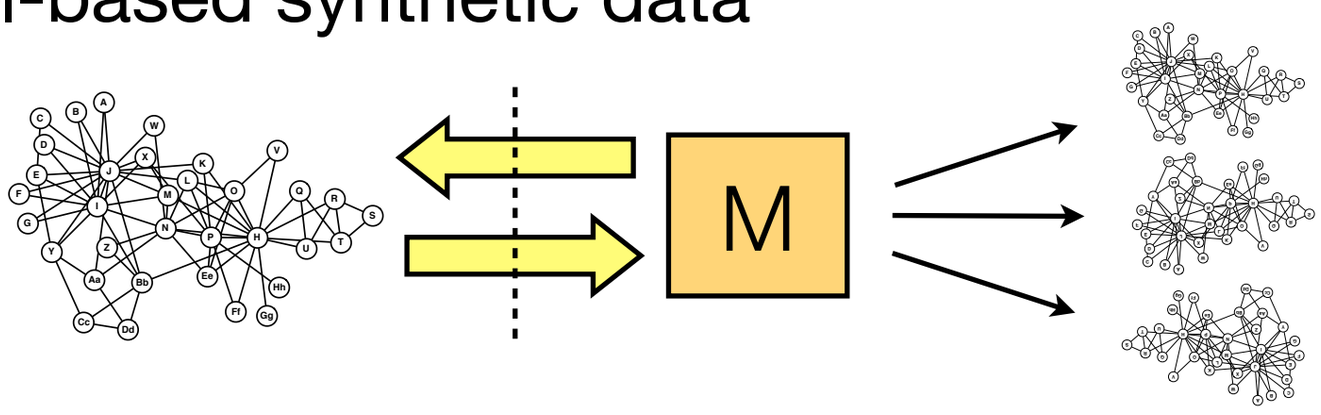
Usability	good
Privacy	no formal guarantees
Accuracy	no formal guarantees
Scalability	sometimes bad

- Output perturbation



Usability	bad for practical analyses
Privacy	formal guarantees
Accuracy	provable bounds
Scalability	very good

- Model-based synthetic data

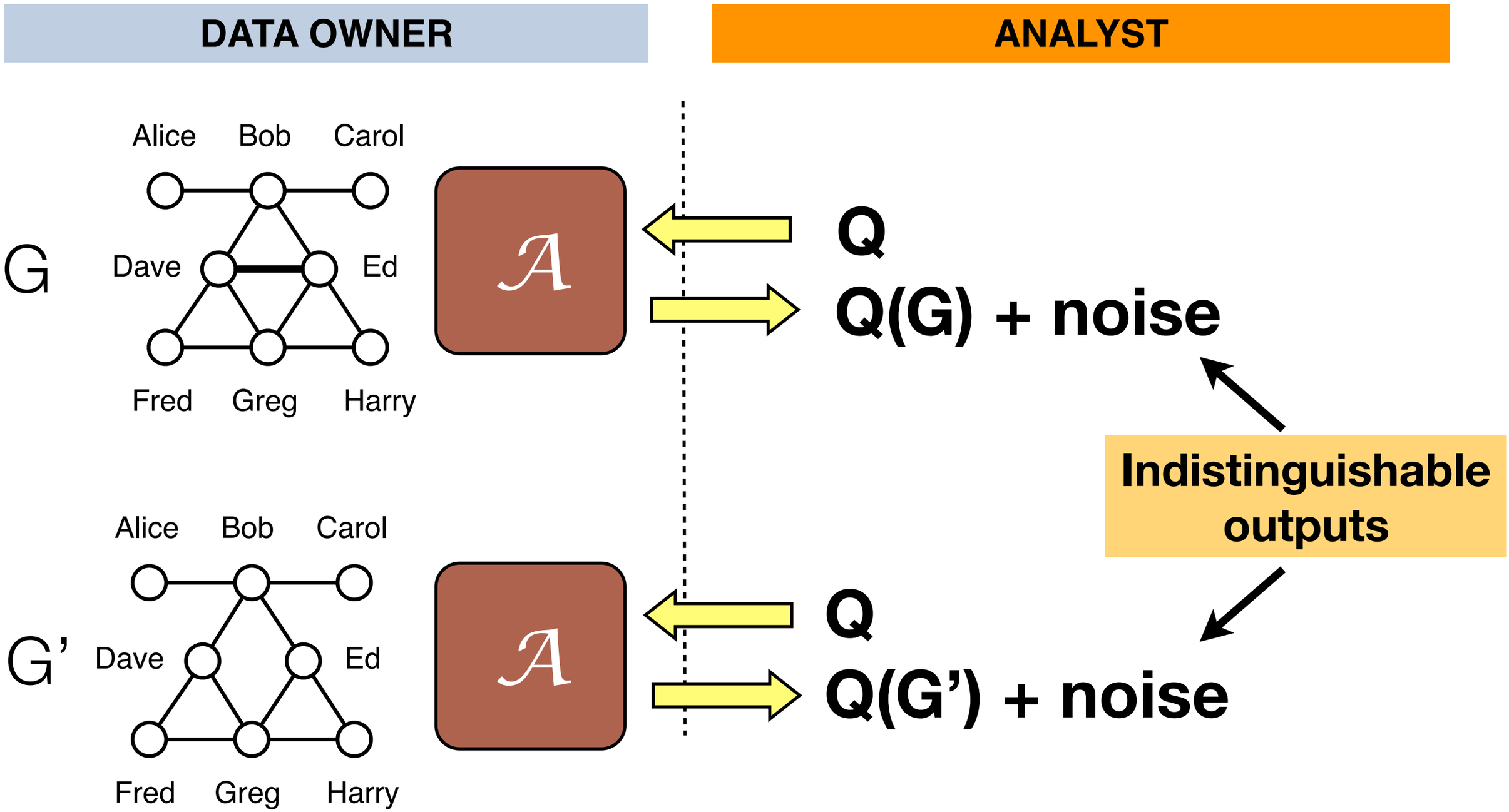


The best of both worlds ??

Outline of the talk

- Differential privacy defined, adapted to graphs
- The accuracy of two topological analyses under differential privacy:
  - Motif analysis
  - **Degree distribution**
- Future goals and open questions

The differential guarantee



Two graphs are **neighbors** if they differ by at most one edge

[Dwork, TCC 06]

Differential privacy

A randomized algorithm A provides **ε-differential privacy** if:
for all neighboring graphs G and G', and
for any set of outputs *S*:

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

epsilon is a privacy parameter

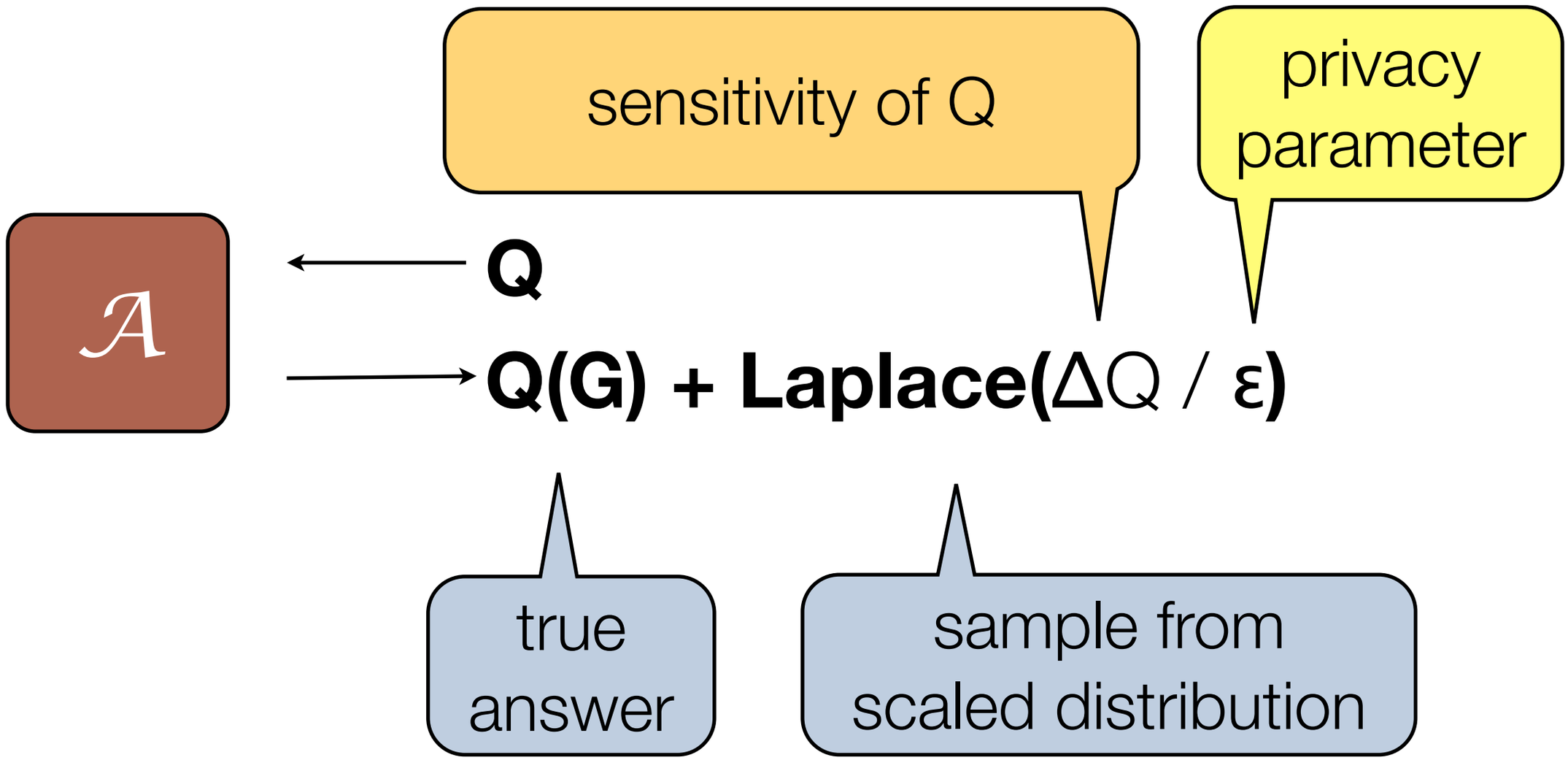
Epsilon is usually small: e.g. if $\epsilon = 0.1$ then $e^{\epsilon} \approx 1.10$

⇩ epsilon = ⇧ stronger privacy

[Dwork, TCC 06]

Calibrating noise

- The following algorithm for answering Q is ε-differentially private:



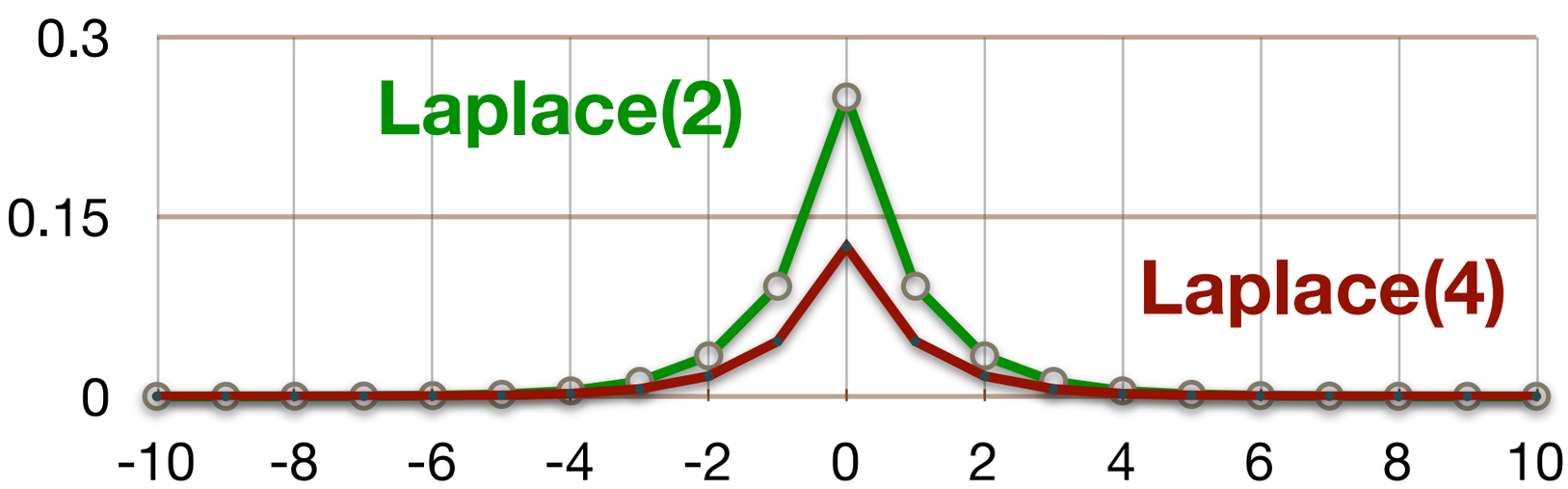
ΔQ: The maximum change in Q, over any two neighboring graphs

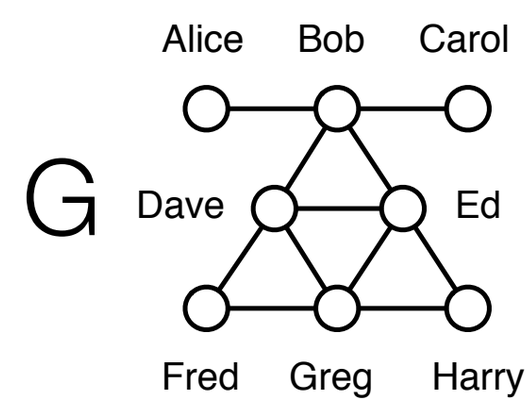
Examples of query sensitivity

query | sensitivity | truth | noise

ε=0.5

Q	ΔQ	Q(G)	Lap(ΔQ / ε)
$\mathbf{deg_A}$ (degree of node A)	1	$\mathbf{deg_{Dave}}$(G) = 4	4+Lap(2)
$\mathbf{cnt_i}$ (# nodes with degree i)	2	$\mathbf{cnt_4}$(G) = 4	4+Lap(4)
$\mathbf{D=[deg_{Dave}, deg_{Ed}, deg_{Fred}]}$	2	**D**(G) = [4,4,2]	[4+Lap(4), 4+Lap(4), 2+Lap(4)]





Differential privacy for networks

A participant's sensitive information is **not** a single edge.

- **edge ε-differential privacy**: algorithm output is largely indistinguishable whether or not any **single edge** is present or absent.
- **k-edge ε-differential privacy**: algorithm output is largely indistinguishable whether or not any **set of k edges** is present or absent.
- **node ε-differential privacy**: algorithm output is largely indistinguishable whether or not any single **node (and all its edges)** is present or absent.

Laplace(ΔQ /ε **)**

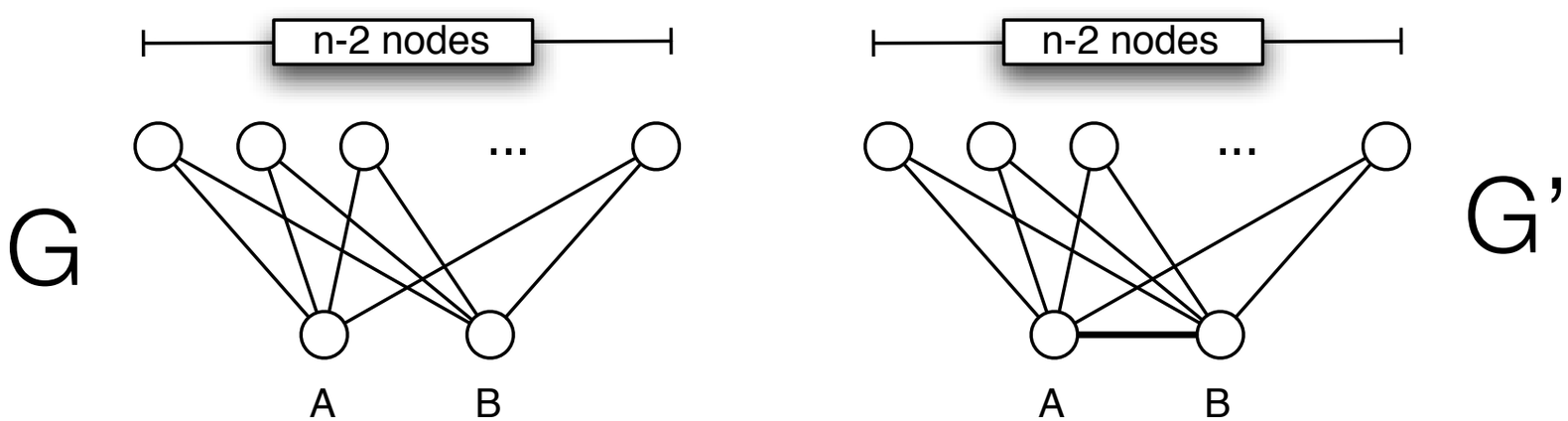
↓

Laplace(ΔQ k /ε **)**

Suppose ΔQ=1. Then Laplace(100) satisfies:
1-edge 0.01-differential privacy
10-edge 0.1-differential privacy

Accurate motif analysis is hard

- Motif analysis measures the frequency of occurrence of small subgraphs in a network.
- Common example: **transitivity** in the network:
  - when A is friends with B and C, are B and C also friends?
  - **$Q_{TRIANGLE}$**: return the number of triangles in the graph



$Q_{TRIANGLE}$ (G) = 0

$Q_{TRIANGLE}$ (G') = n-2

High Sensitivity:

$\Delta Q_{TRIANGLE}=O(n)$

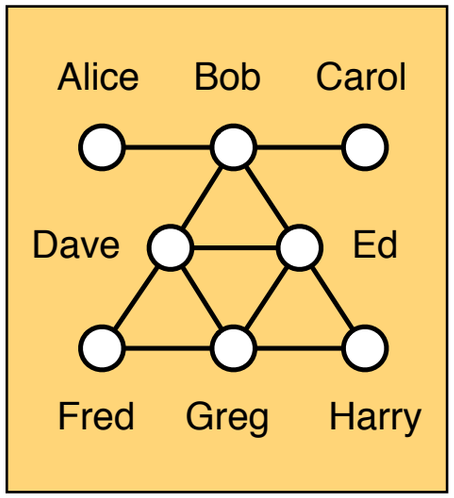
For improved accuracy, see: **[Rastogi, PODS 09]**

[Nissim, STOC 07]

Accurate degree sequence estimation is possible

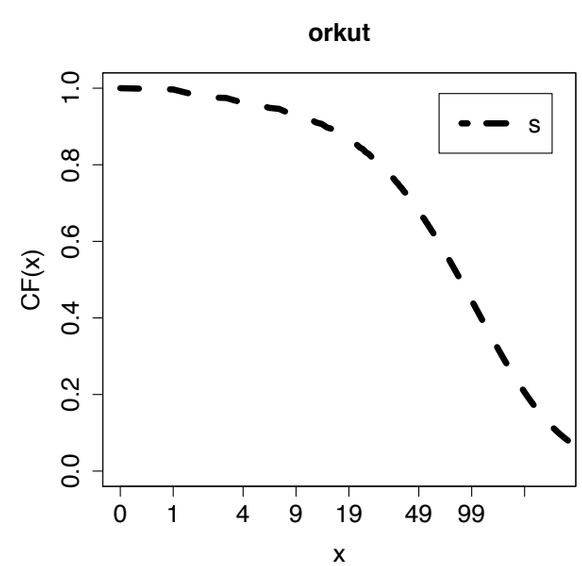
- Degree sequence: the list of degrees of each node in a graph.
- A widely studied property of networks.

G



Orkut crawl

[1,1,2,2,4,4,4,4]
degree sequence of G

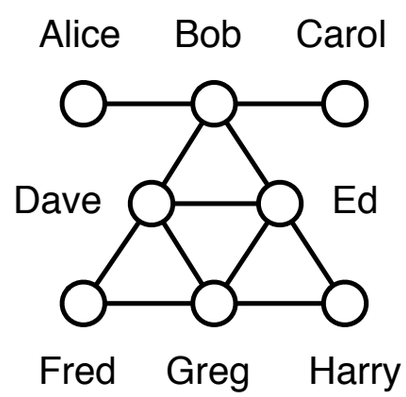


Inverse cummulative distribution

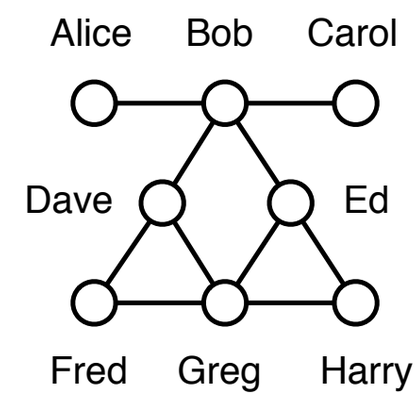
Alternative queries for the degree sequence

count queries		vector queries	
deg_A	return degree of node A	D	$[deg_A, deg_B, \ldots \quad]$
rnk_i	return the rank i^{th} degree	S	$[rnk_1, rnk_2, \ldots \quad cnt_n]$

G

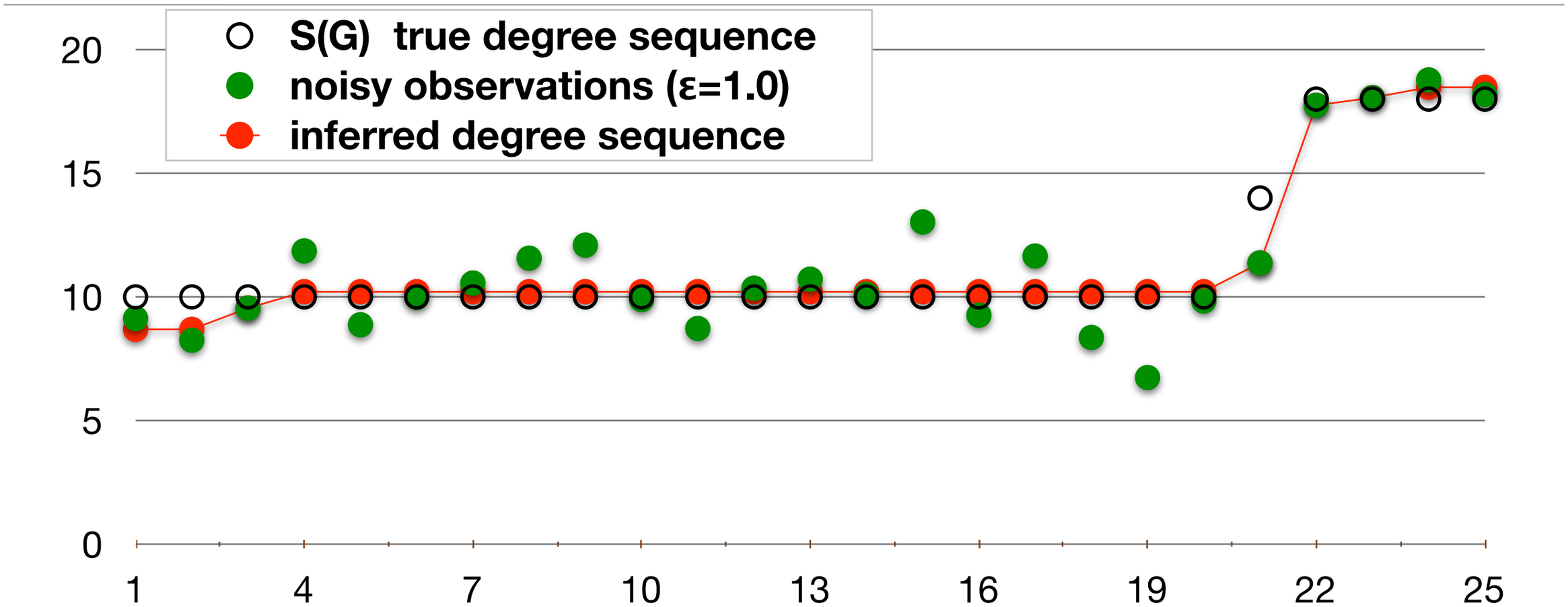


G'



D(G) = [1,4,1,4,4,2,4,2]		D(G') = [1,4,1,<u>3</u>,<u>3</u>,2,4,2]	ΔD=2
S(G) = [1,1,2,2,4,4,4,4]		S(G') = [1,1,2,2,<u>3</u>,<u>3</u>,4,4]	ΔS=2

Using the sort constraint



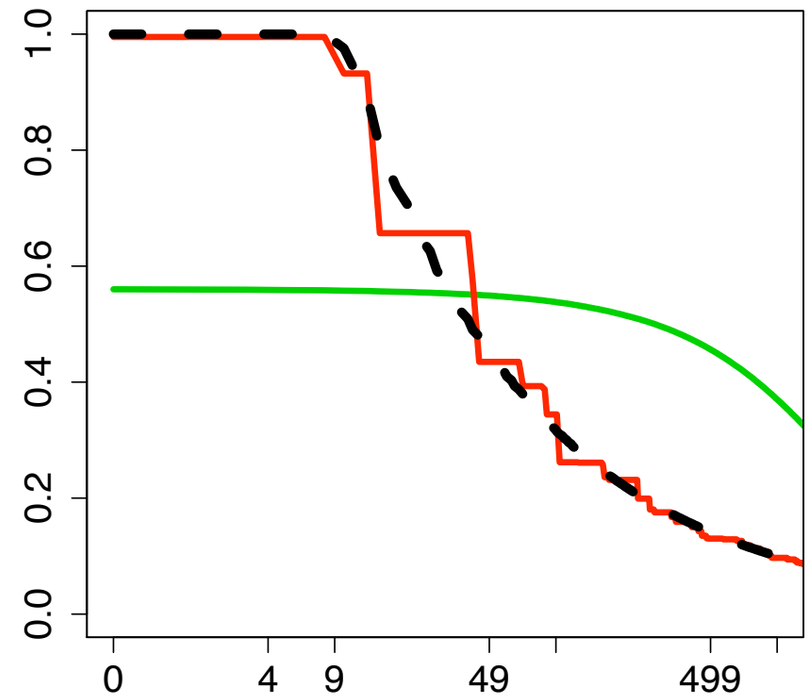
- The output of the sorted degree query is not (in general) sorted.
- We derive a new sequence by computing the **closest** non-decreasing sequence: i.e. minimizing L2 distance.

Experimental results

powerlaw, α=1.5, n=5M

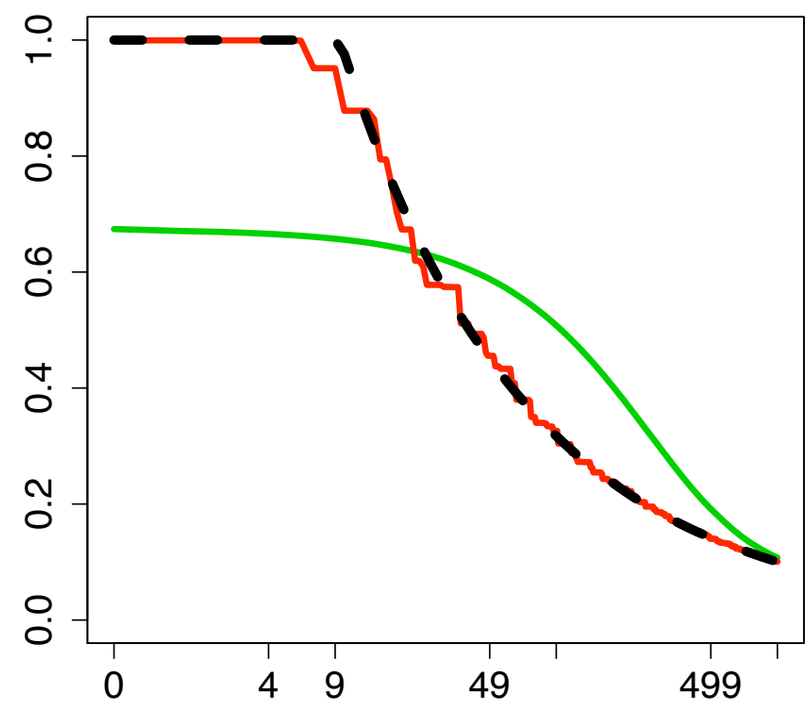
(100-edge, 0.1-differential privacy)

ε=.001

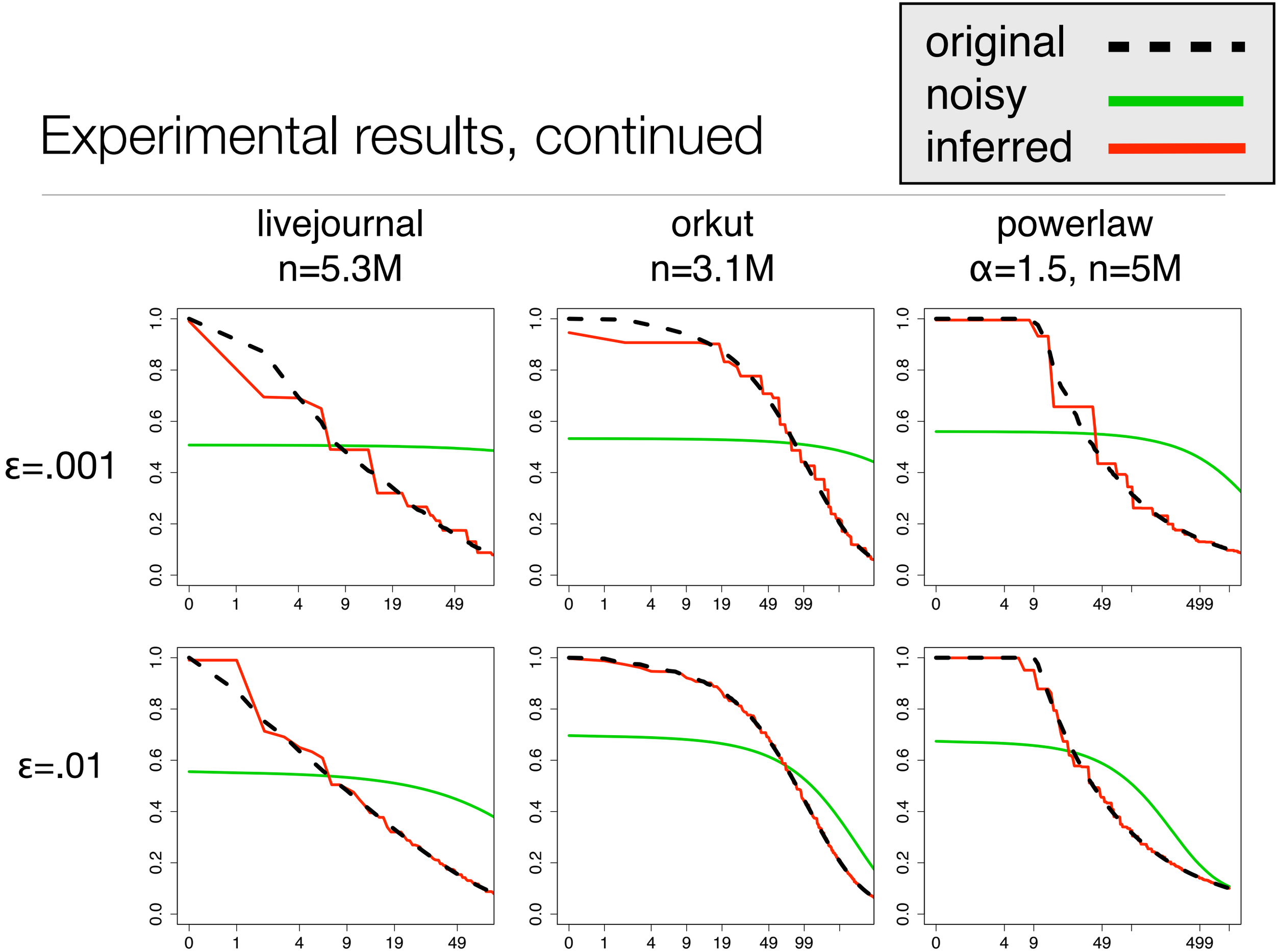


(10-edge, 0.1-differential privacy)

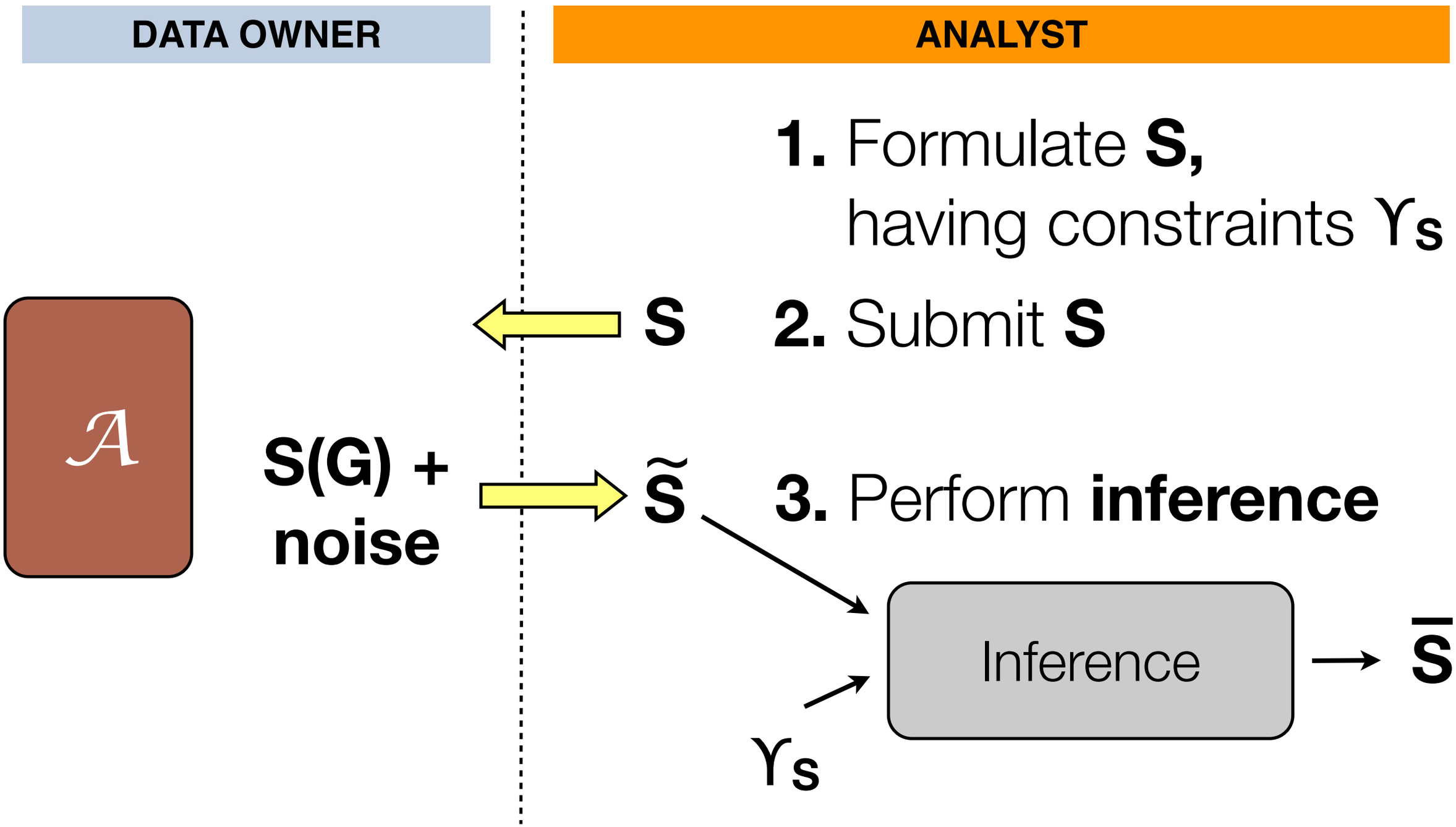
ε=.01



Experimental results, continued



Inference does not weaken privacy



Accuracy is improved without sacrificing privacy!

- Improved accuracy is possible because standard Laplace noise is sufficient (but not necessary) for differential privacy.

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

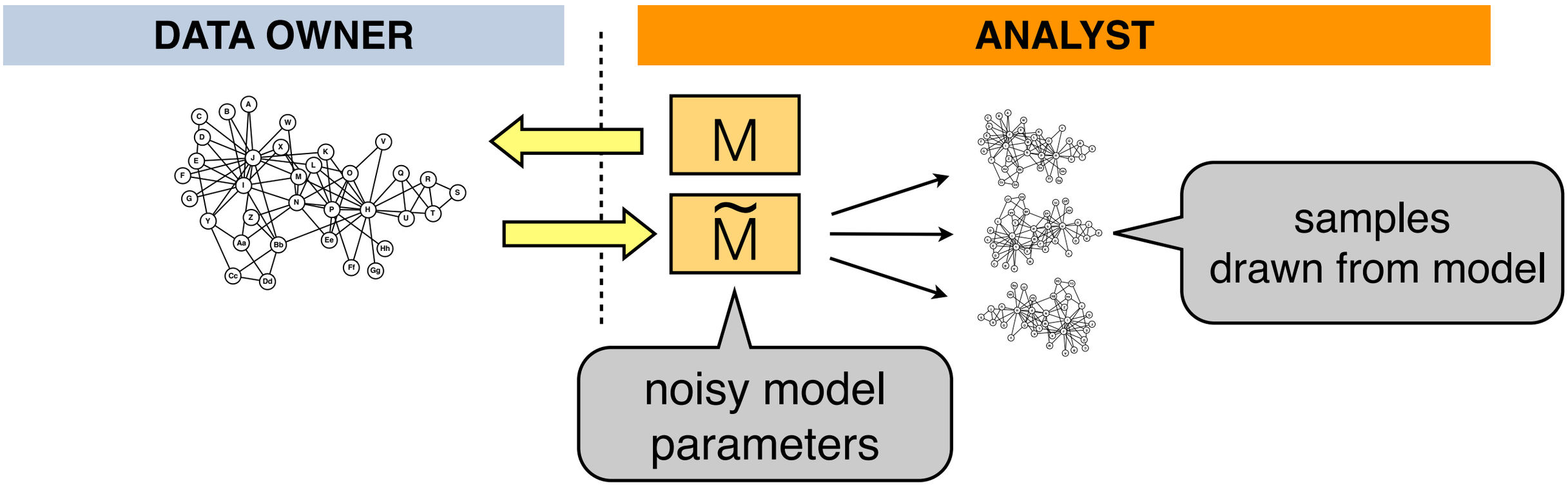
- An immediate consequence: the accuracy achieved **depends on the input sequence**.

Mean squared error, $\tilde{\mathbf{S}}$ $\Theta(n/\epsilon^2)$

Mean squared error, $\bar{\mathbf{S}}$ $O(d\log^3 n/\epsilon^2)$

number of distinct degrees

Toward differentially-private synthetic data



- To realize the benefits of synthetic data, data owner can release noisy parameters of network model.
- Baseline: the degree distribution as network model
  - Deriving the power law parameter — very accurate
  - Measuring clustering coefficient — not constrained by deg. distr.

A few open questions

- For graph measure X, how accurately can we estimate X under differential privacy?
- Is it socially acceptable to offer weaker privacy protection to high-degree nodes (as in k-edge differential privacy)?
- Can we generate accurate synthetic networks under differential privacy?
  - Can the analyst explore an approximate synthetic network, then query the original network under output perturbation?

Questions?

Additional details on our work may be found here:

- **[Hay, ICDM 09]** M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In International Conference on Data Mining (ICDM), To appear, 2009.

- **[Hay, CoRR 09]** M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially-private queries through consistency. CoRR, abs/0904.0942, 2009. (under review)

- **[Rastogi, PODS 09]** V. Rastogi, M. Hay, G. Miklau, and D. Suciu. Relationship privacy: Output perturbation for queries with joins. In Principles of Database Systems (PODS), 2009.

- **[Hay, VLDB 08]** M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis. Resisting structural identification in anonymized social networks. In VLDB Conference, 2008.

References

- **[Liu, SIGMOD 08]** K. Liu and E. Terzi. Towards identity anonymization on graphs. In SIGMOD, 2008.
- **[Zhou, ICDE 08]** B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In ICDE, 2008.
- **[Zou, VLDB 09]** L. Zou, L. Chen, and T. Ozsu. K-automorphism: A general framework for privacy preserving network publication. In Proceedings of VLDB Conference, 2009.
- **[Ying, SDM 2008]** X. Ying and X. Wu. Randomizing social networks: a spectrum preserving approach. In SIAM International Conference on Data Mining, 2008.
- **[Cormode, VLDB 08]** G. Cormode, D. Srivastava, T. Yu, and Q. Zhang. Anonymizing bipartite graph data using safe groupings. In VLDB Conference, 2008.

References (con't)

- **[Cormode, VLDB 09]** G. Cormode, D. Srivastava, S. Bhagat, and B. Krishnamurthy. Class-based graph anonymization for social network data. In VLDB Conference, 2009.

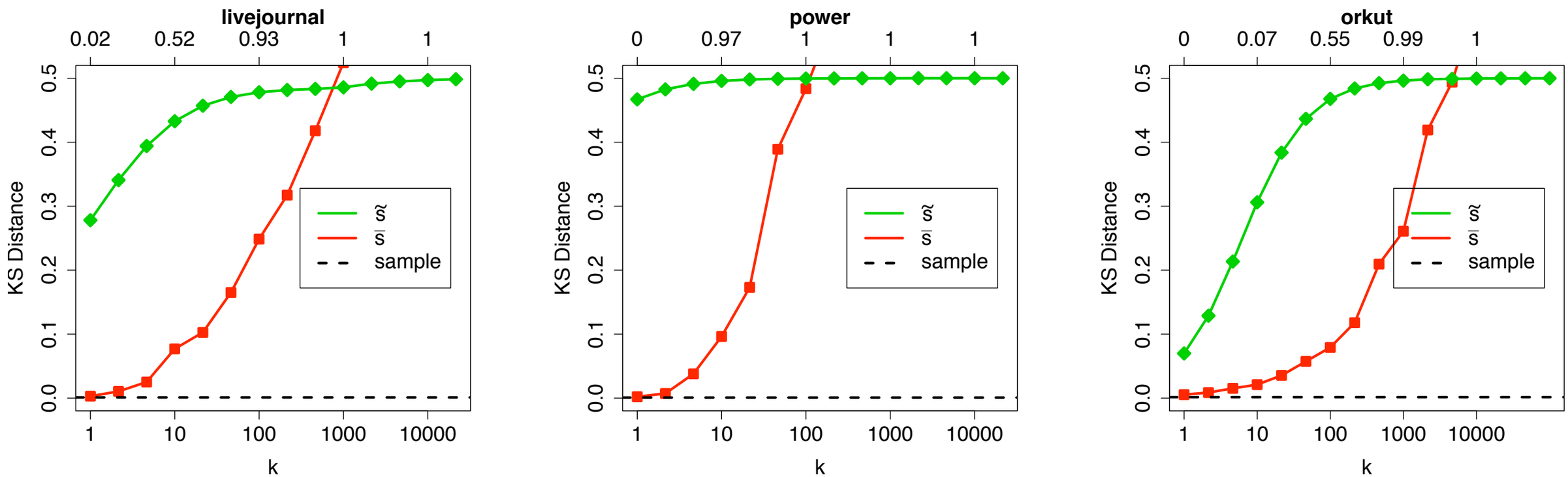
- **[Campan, PinKDD 08]** A. Campan and T. M. Truta. A clustering approach for data and structural anonymity in social networks. In PinKDD, 2008.

- **[Rastogi, VLDB 07]** V. Rastogi, S. Hong, and D. Suciu. The boundary between privacy and utility in data publishing. In VLDB, pages 531–542, 2007.

- **[Dwork, TCC 06]** C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In Third Theory of Cryptography Conference, 2006.

- **[Nissim, STOC 07]** K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In STOC, pages 75–84, 2007.

THEOREM 1. *Denote* $L_k = \min_{j \in [k,n]} \max_{i \in [1,j]} M[i,j]$ *and* $U_k = \max_{i \in [1,k]} \min_{j \in [i,n]} M[i,j]$. *The minimum* L_2 *solution* $\overline{s} = \text{MIN} L_2(\tilde{s}, \gamma_{\mathbf{S}})$, *is unique and given by:* $\overline{s}[k] = L_k = U_k$.


livejournal
0.02
0.52
0.93
1
1
KS Distance
0.0
0.1
0.2
0.3
0.4
0.5
1
10
100
1000
10000
k
sl
s
sample
power
0
0.97
1
1
1
KS Distance
0.0
0.1
0.2
0.3
0.4
0.5
1
10
100
1000
10000
k
sl
s
sample
orkut
0
0.07
0.55
0.99
1
KS Distance
0.0
0.1
0.2
0.3
0.4
0.5
1
10
100
1000
10000
k
sl
s
sample