

Enabling Accurate Analysis of Private Network Data

Michael Hay

Joint work with

Gerome Miklau, David Jensen, Chao Li, Don Towsley

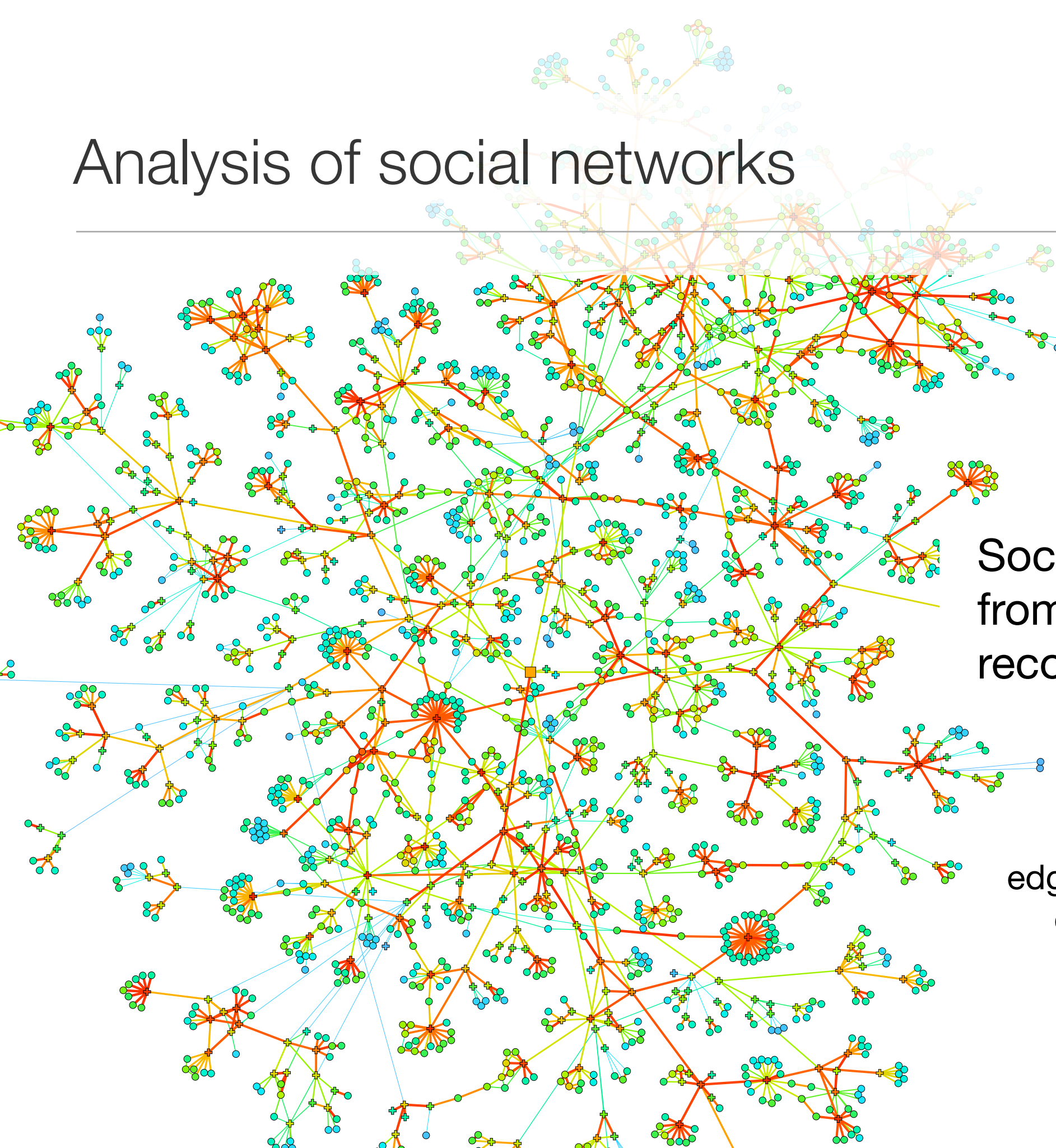
University of Massachusetts, Amherst

Vibhor Rastogi, Dan Suciu

University of Washington

October 8, 2009

Analysis of social networks



Social network derived
from mobile phone call
records [Onnela, PNAS 07]

4.6M nodes
7.0M edges

edge if reciprocal phone calls
during 18 week interval

Analysis of social networks

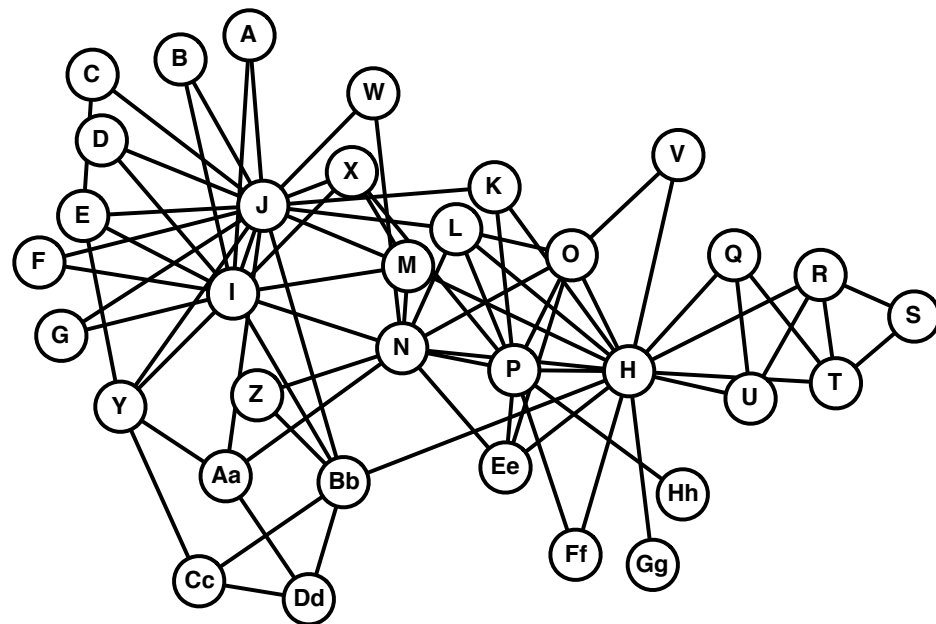


Social network derived from mobile phone call records [Onnela, PNAS 07]

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

How to achieve both privacy and utility?

DATA OWNER



Private network

ANALYST

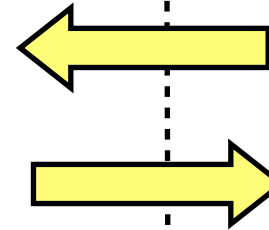
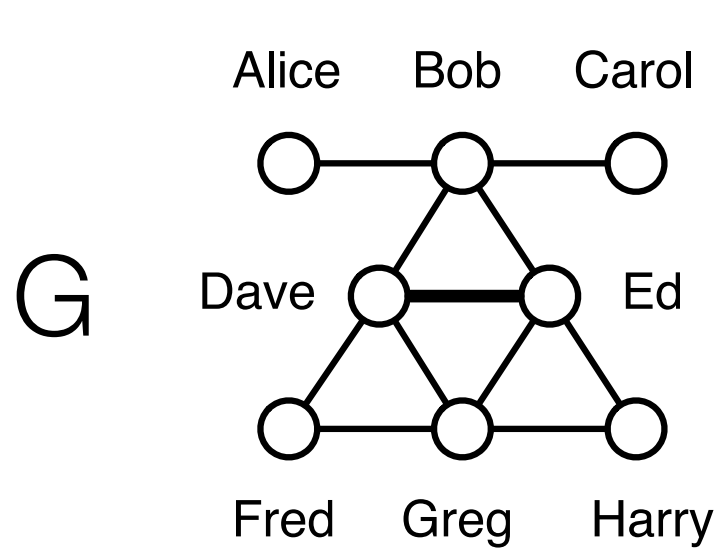
- Q** *what is diameter?*
- Q** *what is maximum degree?*
- Q** *how many 3 cliques?*
- Q** *is Alice ~~connected~~ to Bob?*

Allow aggregate statistics
provided facts about individuals are not disclosed

Query answer perturbation

DATA OWNER

ANALYST



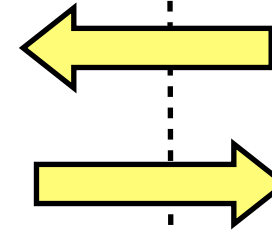
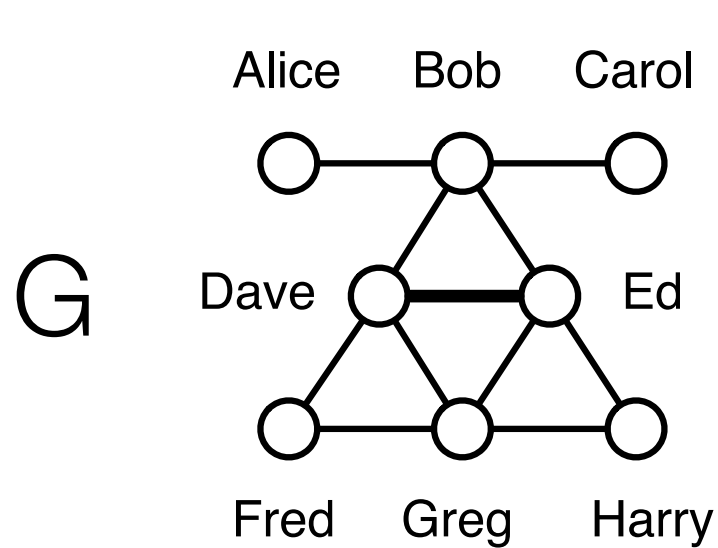
Q

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

Query answer perturbation

DATA OWNER

ANALYST



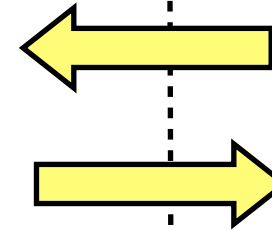
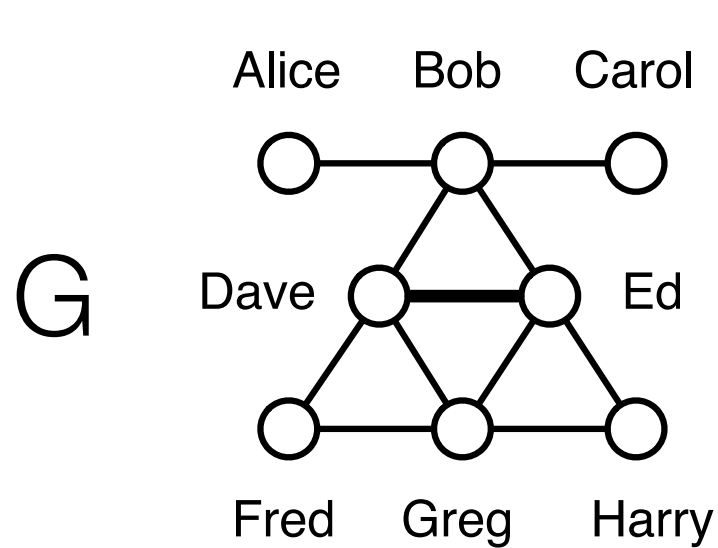
Q *how many edges?*

A = Q(G) + noise

Query answer perturbation

DATA OWNER

ANALYST



Q *how many edges?*

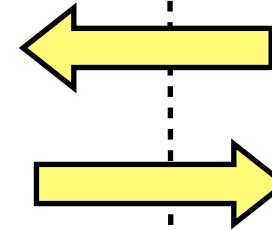
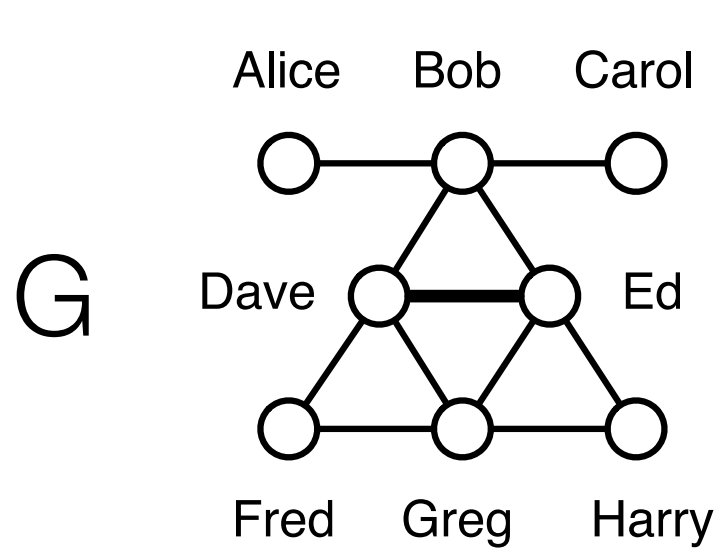
A = Q(G) + noise

11

Query answer perturbation

DATA OWNER

ANALYST



Q *how many edges?*

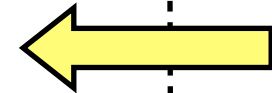
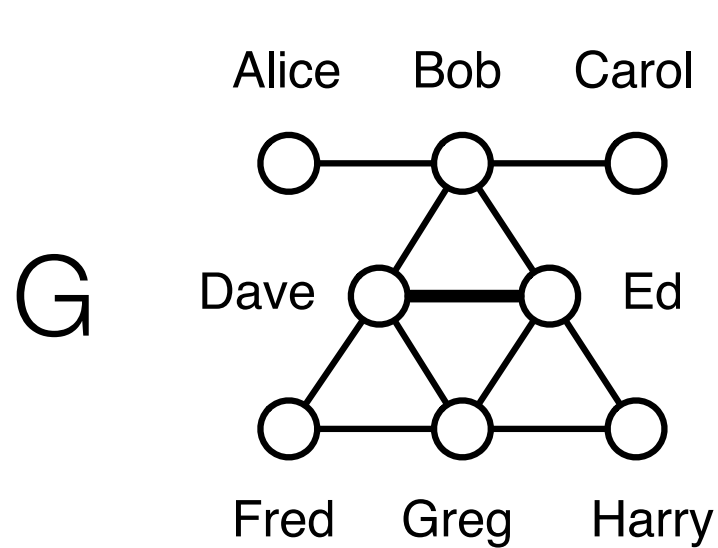
A = Q(G) + noise

11 + 2.3

Query answer perturbation

DATA OWNER

ANALYST



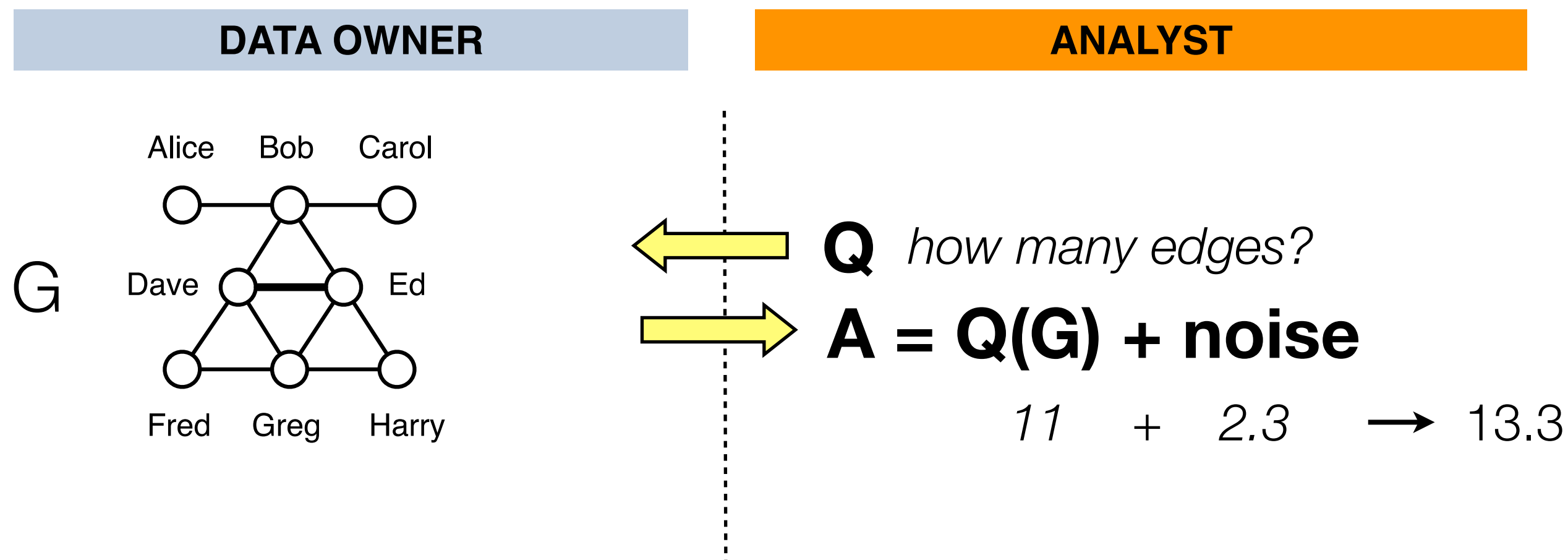
Q *how many edges?*



A = Q(G) + noise

11 + 2.3 → 13.3

Query answer perturbation

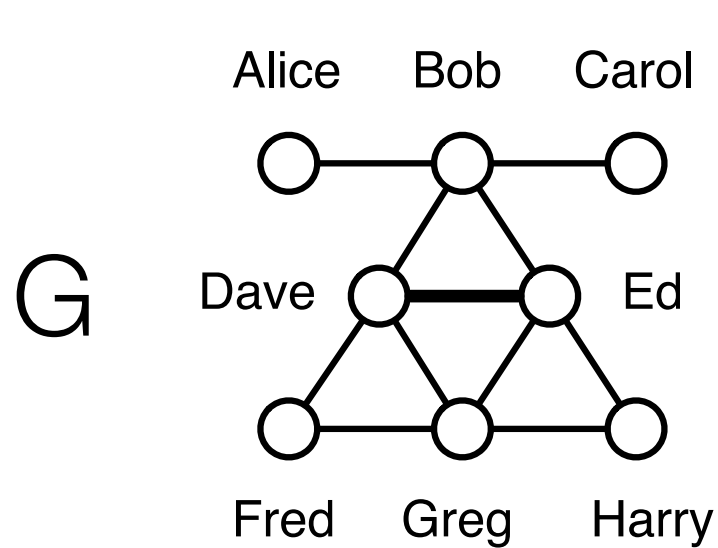


- Dwork, McSherry, Nissim, Smith [Dwork, TCC 06] have described an answer perturbation mechanism satisfying ***differential privacy***.
- Comparatively few results for these techniques applied to graphs.

Query answer perturbation

DATA OWNER

ANALYST



Q *how many edges?*



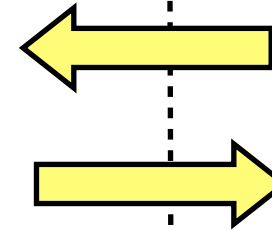
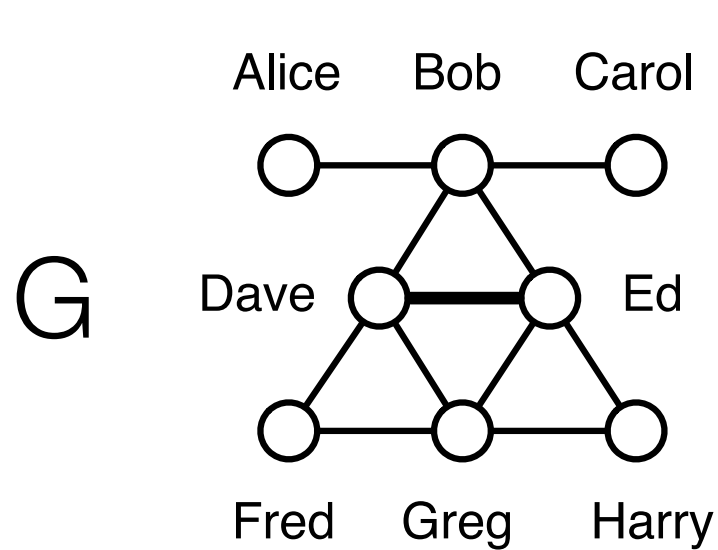
A = Q(G) + noise

11 + 2.3 → 13.3

Query answer perturbation

DATA OWNER

ANALYST



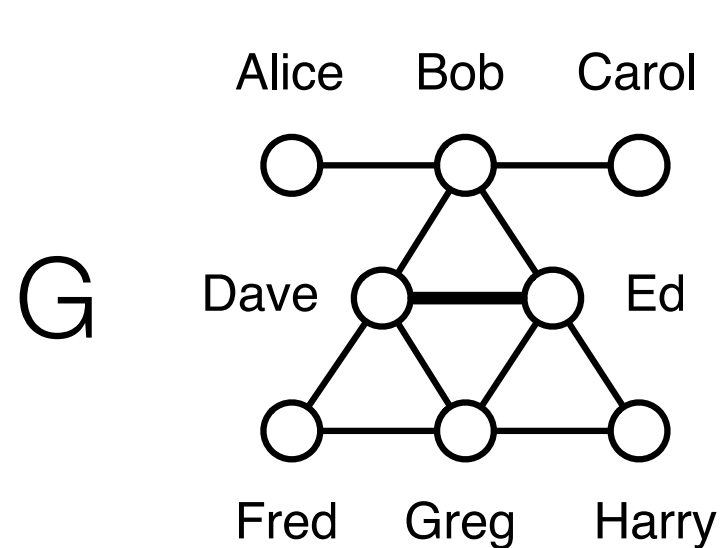
Q *how many edges?*

A = Q(G) + noise

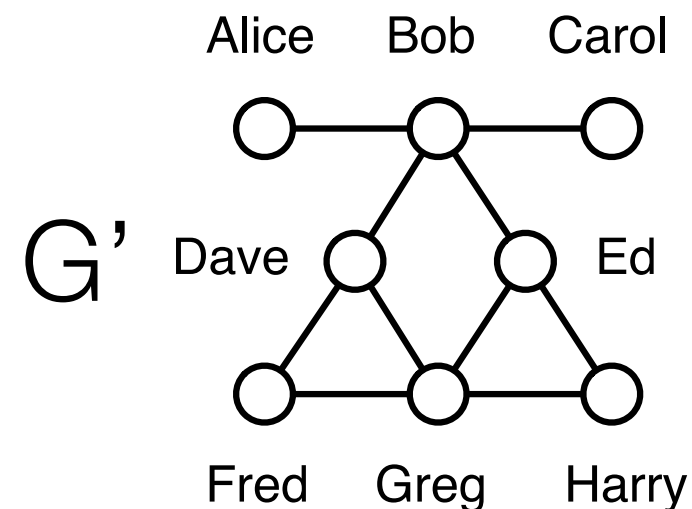
Query answer perturbation

DATA OWNER

ANALYST



← Q *how many edges?*
 → $A = Q(G) + \text{noise}$

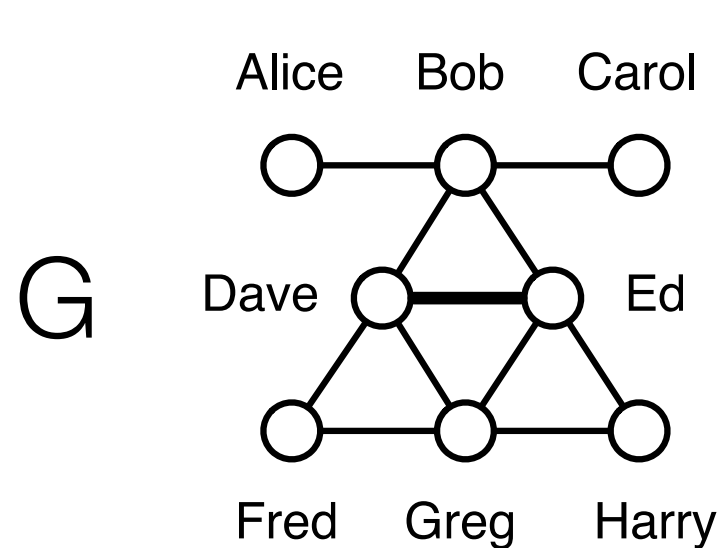


← Q
 → $A = Q(G') + \text{noise}$

Query answer perturbation

DATA OWNER

ANALYST

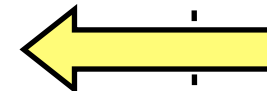
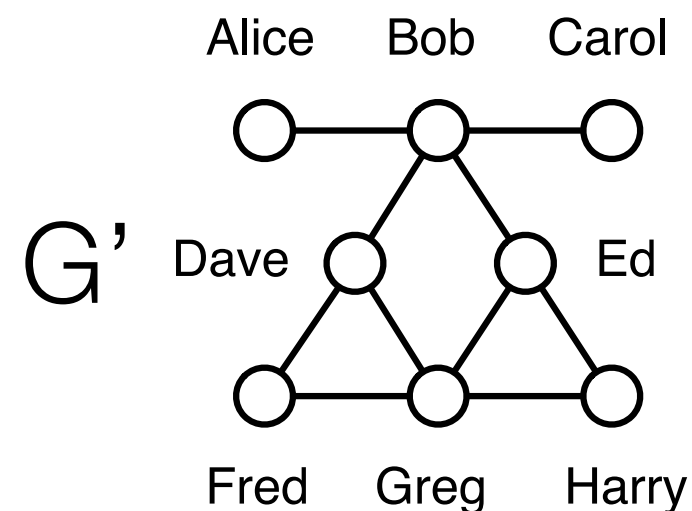


Q *how many edges?*

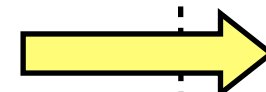


$A = Q(G) + \text{noise}$

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



Q

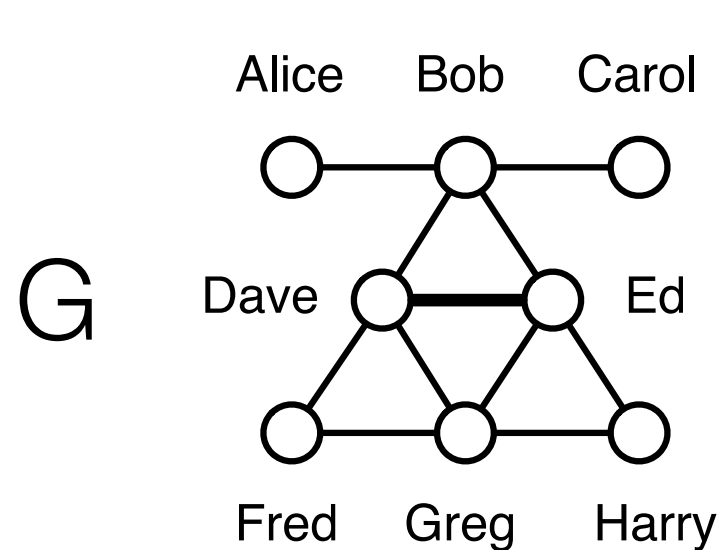


$A = Q(G') + \text{noise}$

Query answer perturbation

DATA OWNER

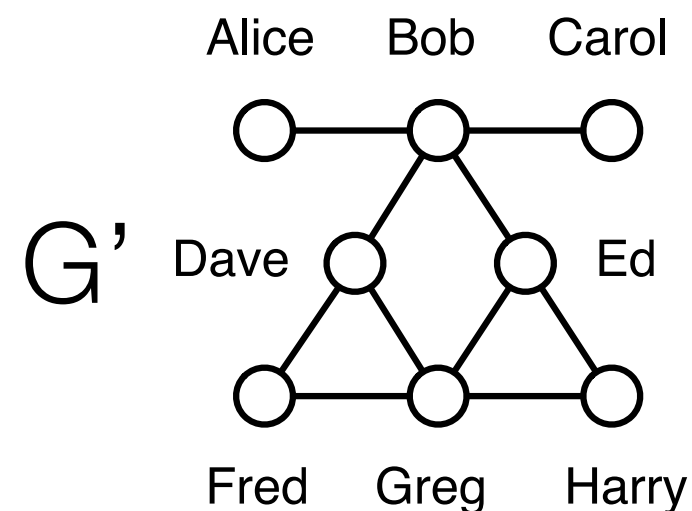
ANALYST



$\leftarrow Q$ *how many edges?*

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

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



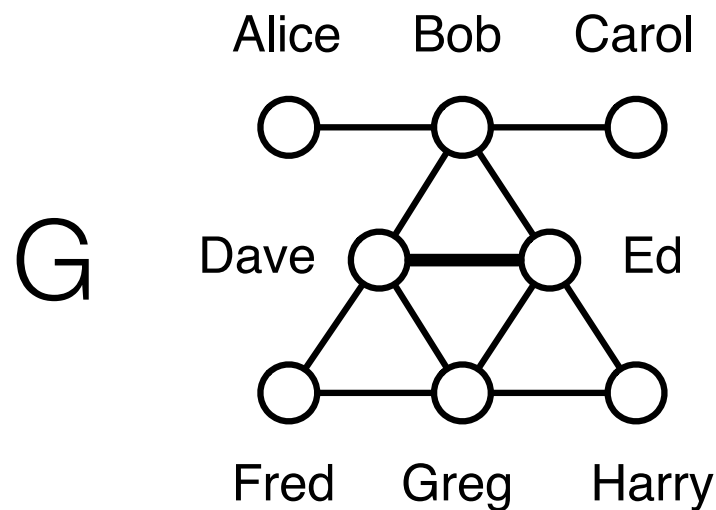
$\leftarrow Q$

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

Query answer perturbation

DATA OWNER

ANALYST

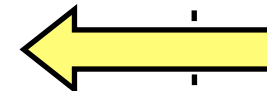


Q *how many edges?*

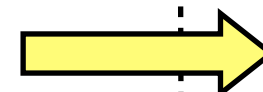


$A = Q(G) + \text{noise}$

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

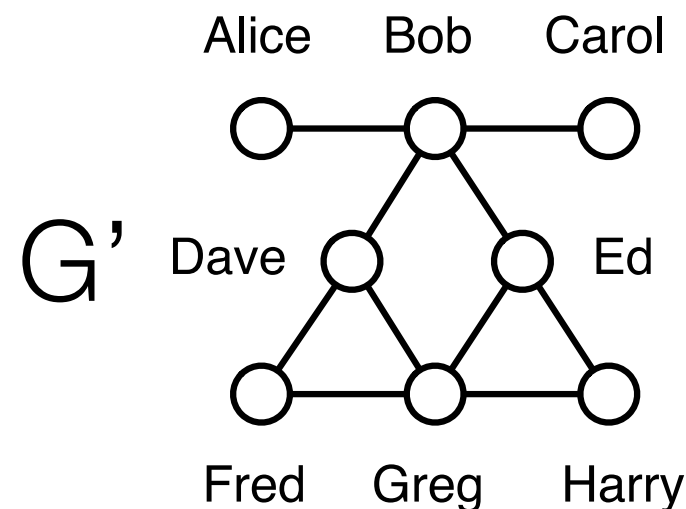


Q



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

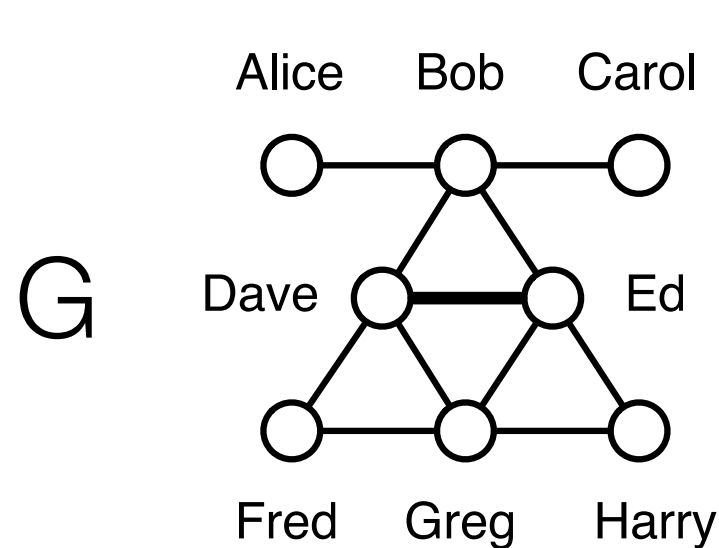
$$\Pr[A = x \mid \mu = Q(G')]$$



Query answer perturbation

DATA OWNER

ANALYST

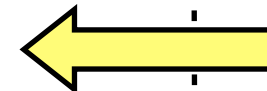
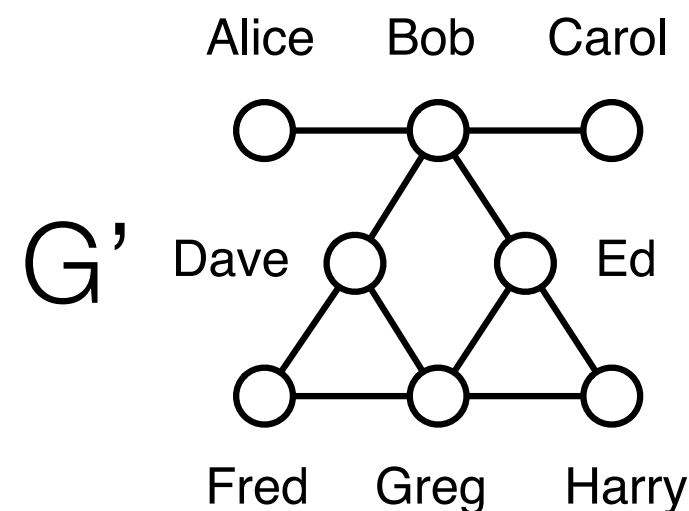


Q *how many edges?*

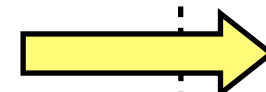


$A = Q(G) + \text{noise}$

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



Q



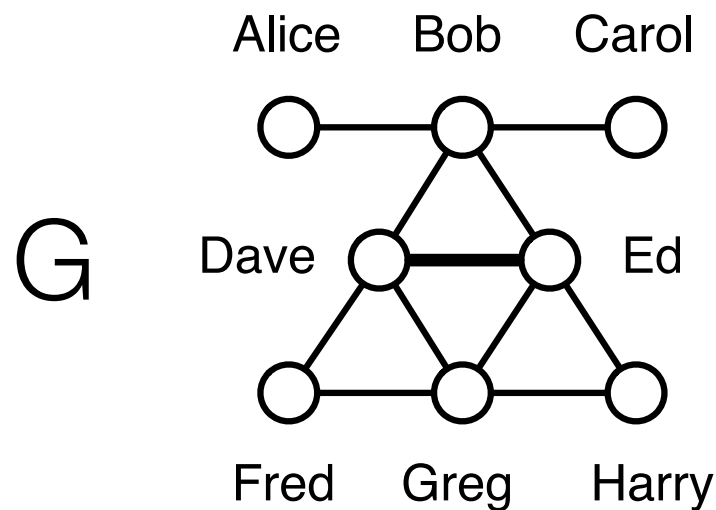
$A = Q(G') + \text{noise}$

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

Query answer perturbation

DATA OWNER

ANALYST

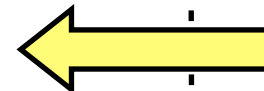


Q *how many edges?*



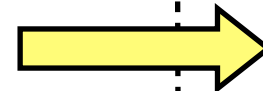
$A = Q(G) + \text{noise}$

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



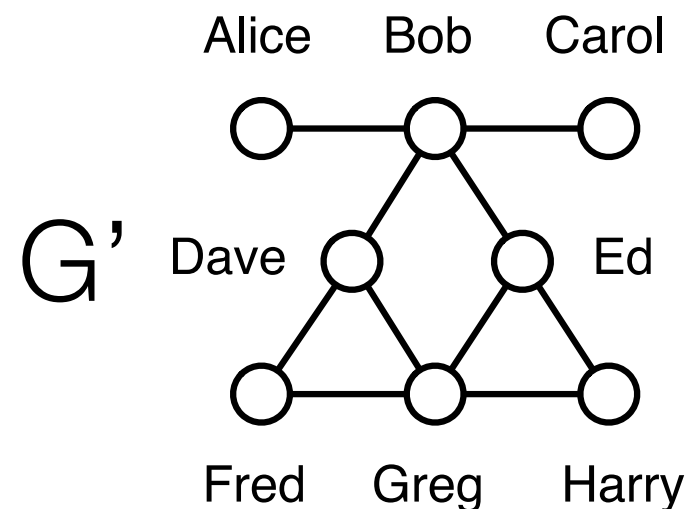
Q

differ by at most
factor of e^ϵ



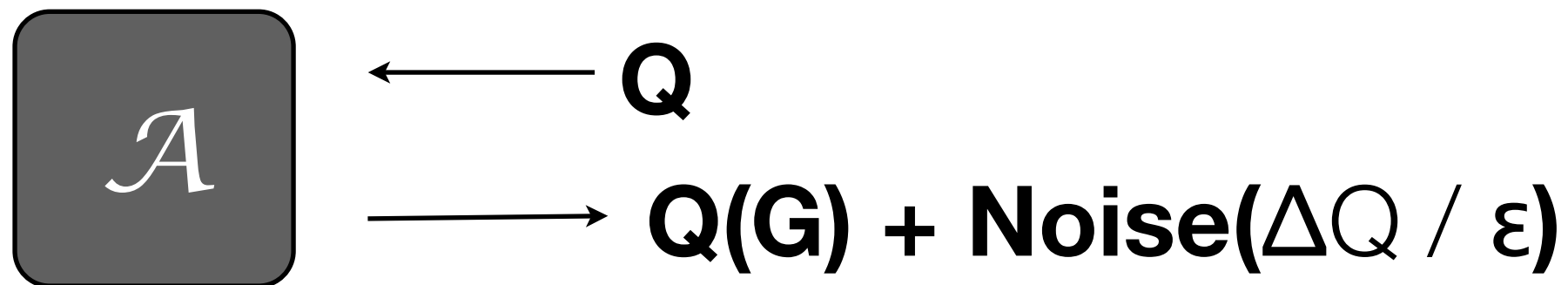
$A = Q(G') + \text{noise}$

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



Calibrating noise

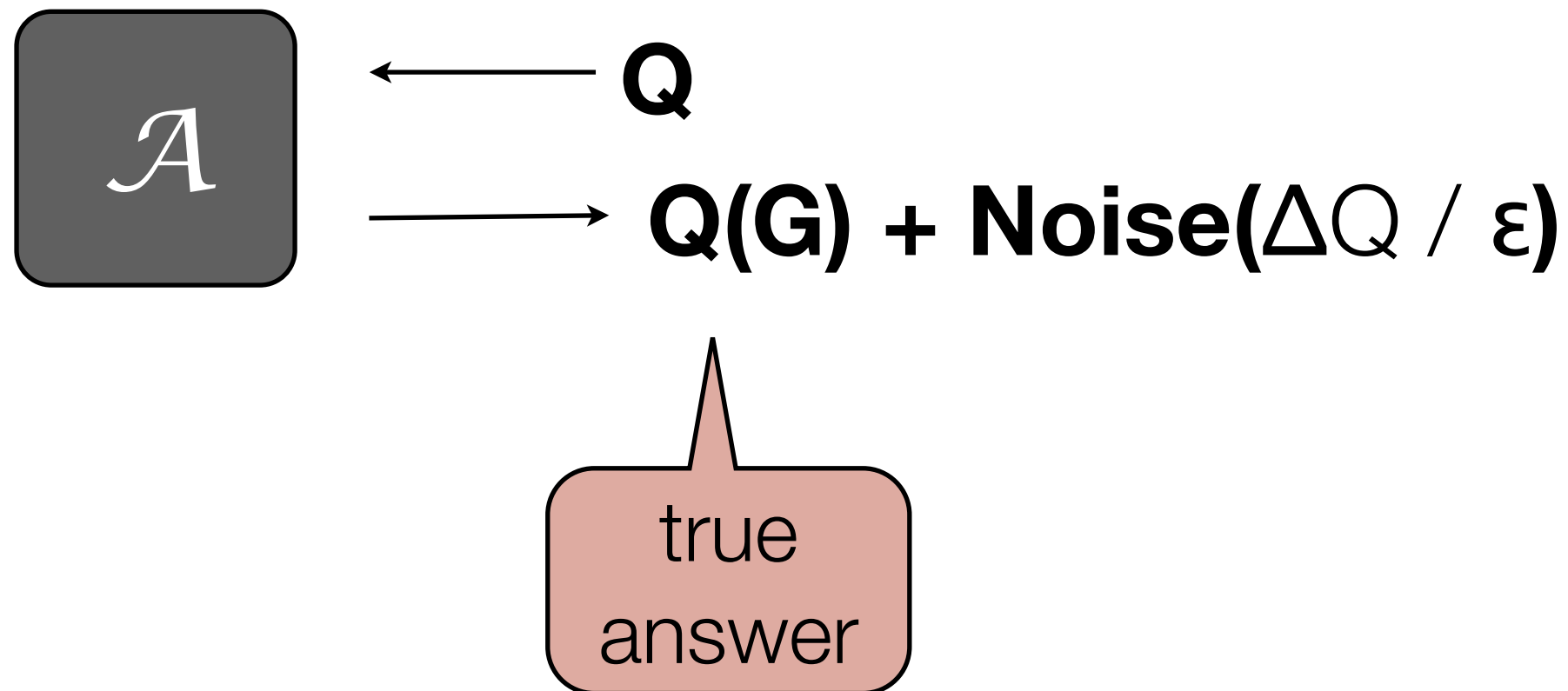
- The following algorithm for answering Q is ϵ -differentially private:



ΔQ : Max change in Q , over any two graphs differing by single edge

Calibrating noise

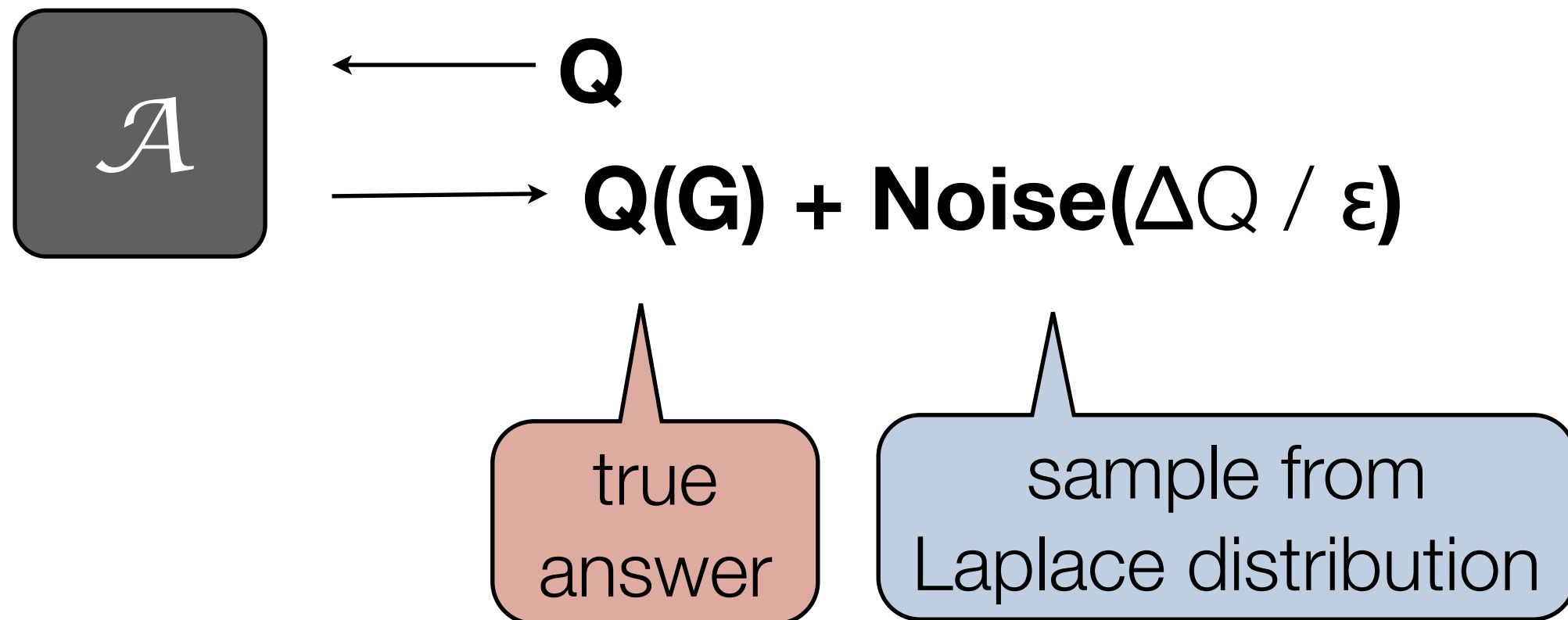
- The following algorithm for answering Q is ϵ -differentially private:



ΔQ : Max change in Q , over any two graphs differing by single edge

Calibrating noise

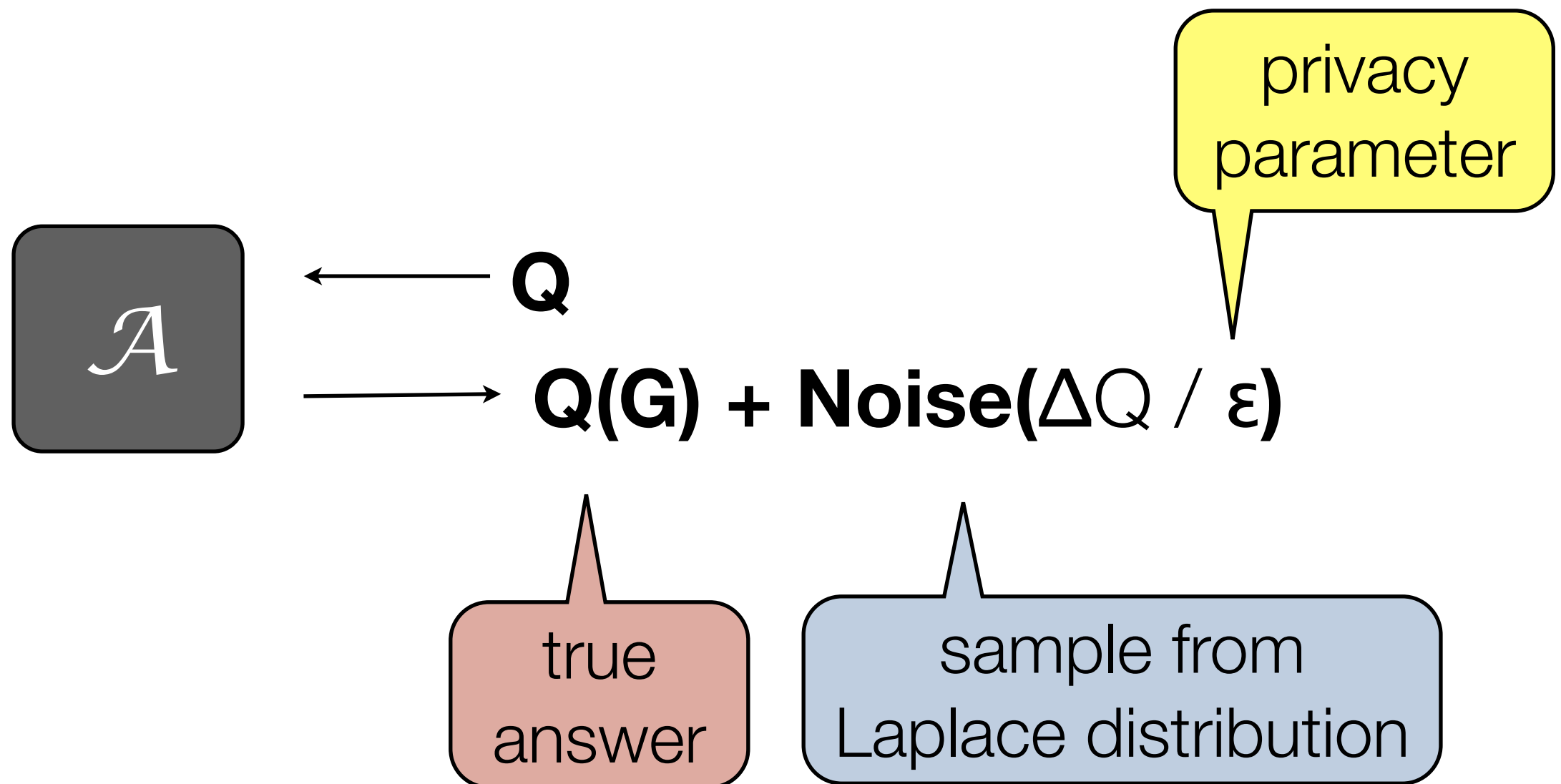
- The following algorithm for answering Q is ϵ -differentially private:



ΔQ : Max change in Q , over any two graphs differing by single edge

Calibrating noise

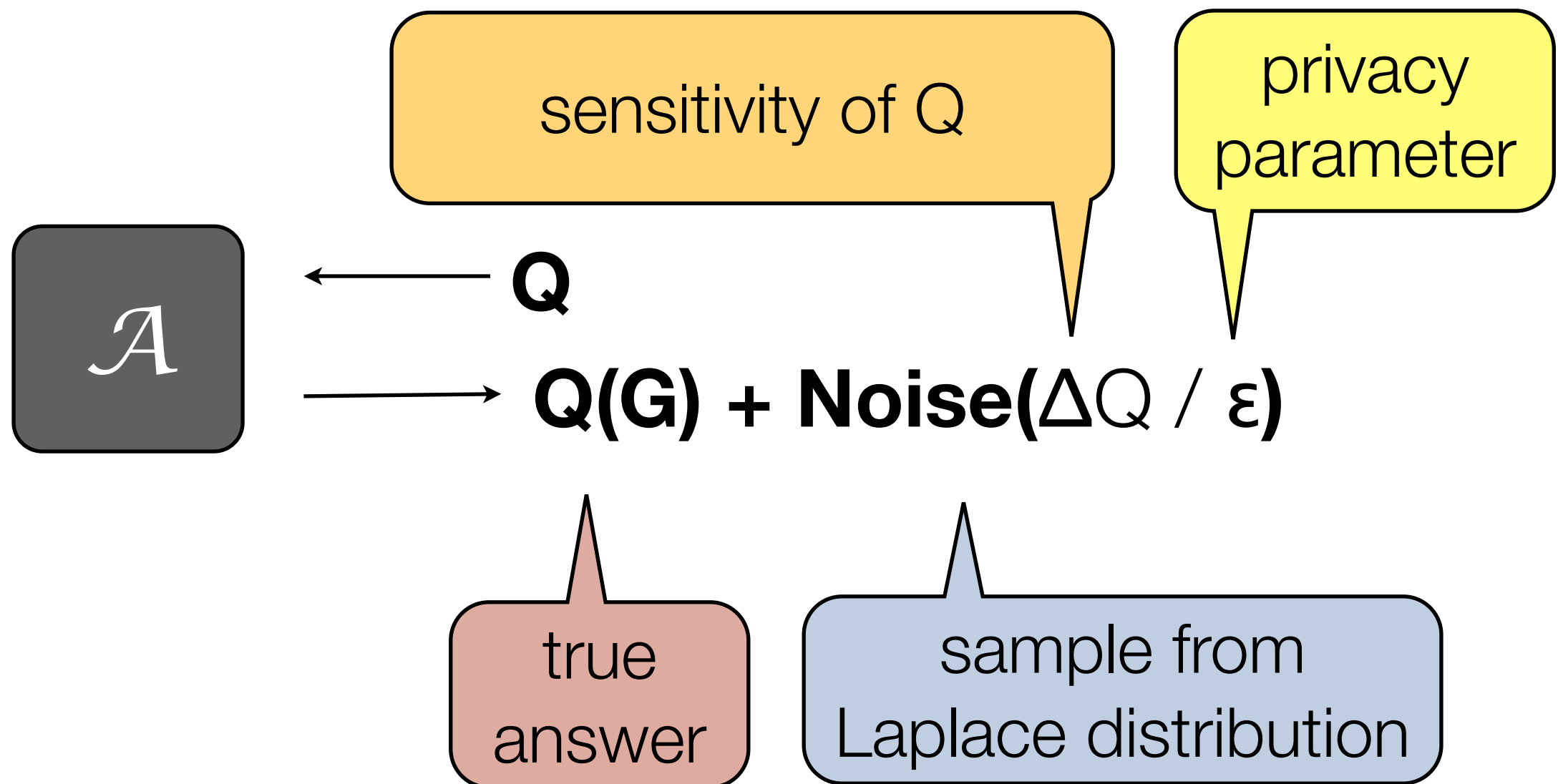
- The following algorithm for answering Q is ϵ -differentially private:



ΔQ : Max change in Q , over any two graphs differing by single edge

Calibrating noise

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



ΔQ : Max change in Q , over any two graphs differing by single edge

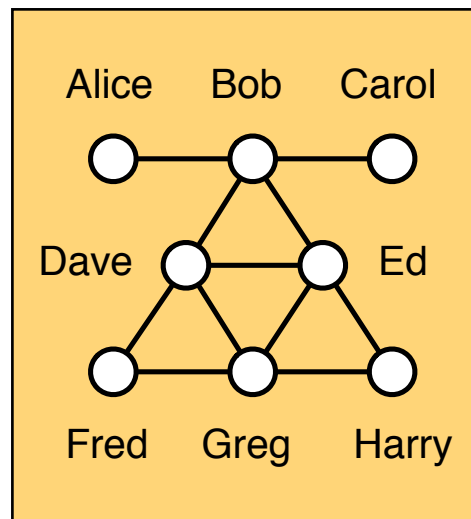
Positive results in differential privacy

- 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?

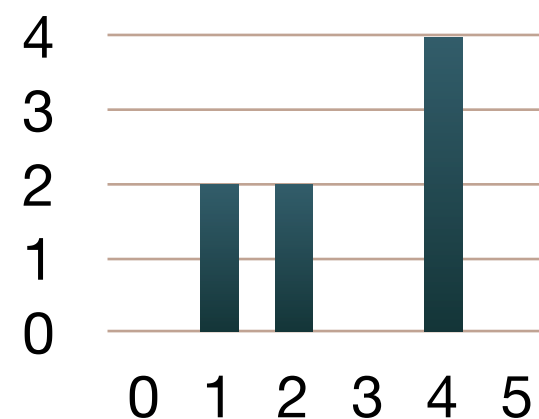
Accurate degree distribution estimation is possible

- Degree distribution: the frequency of each degree in graph.
- A widely studied property of networks.

G

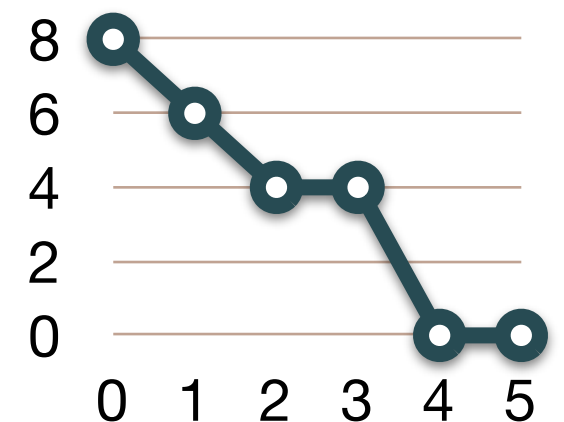


Histogram

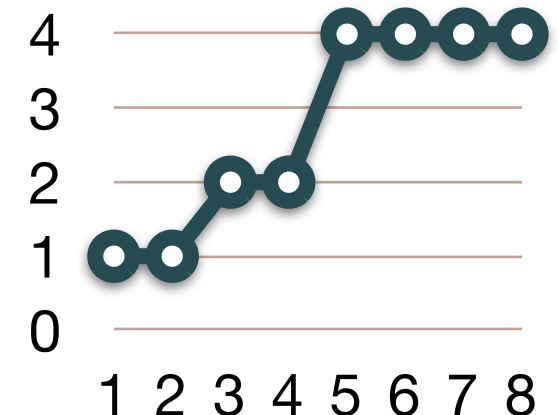
Degree sequence
as a vector

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

CCDF



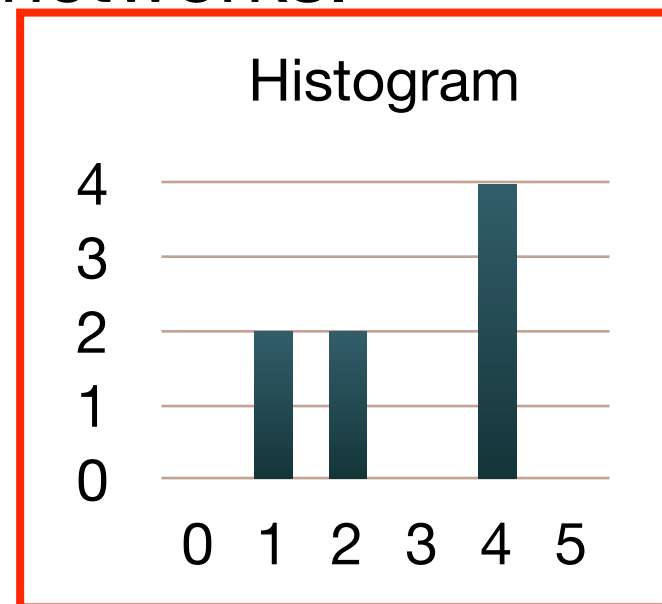
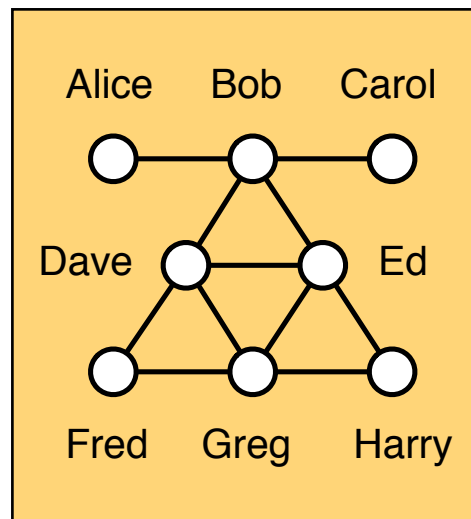
Degree sequence



Accurate degree distribution estimation is possible

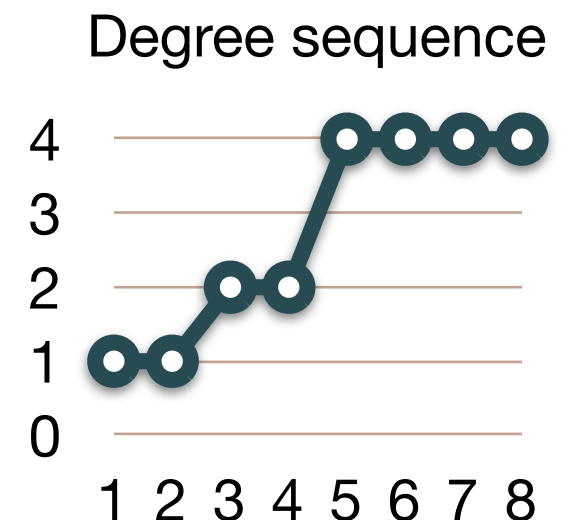
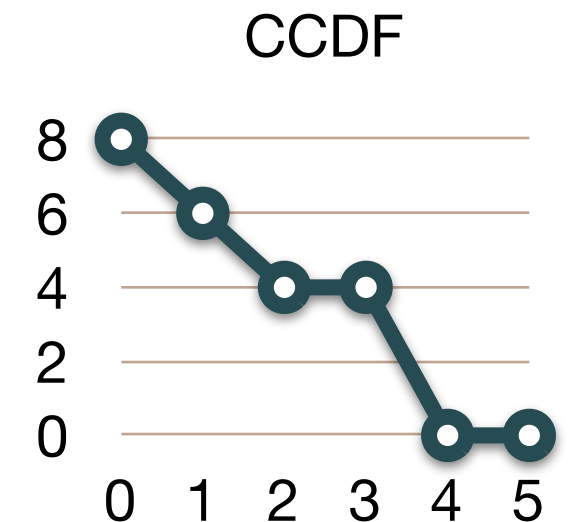
- Degree distribution: the frequency of each degree in graph.
- A widely studied property of networks.

G



Degree sequence
as a vector

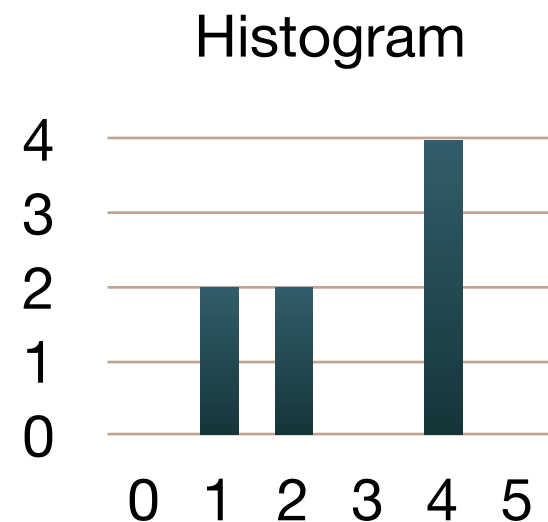
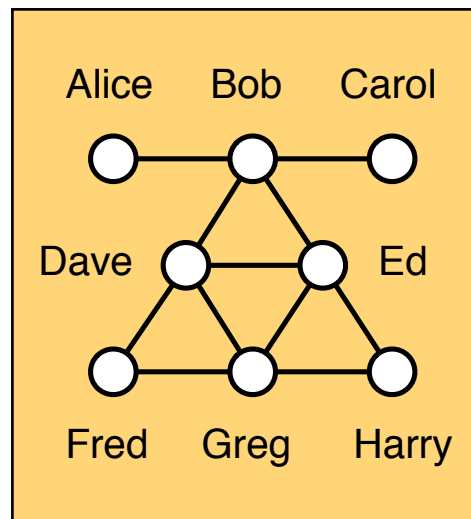
[1,1,2,2,4,4,4,4]



Accurate degree distribution estimation is possible

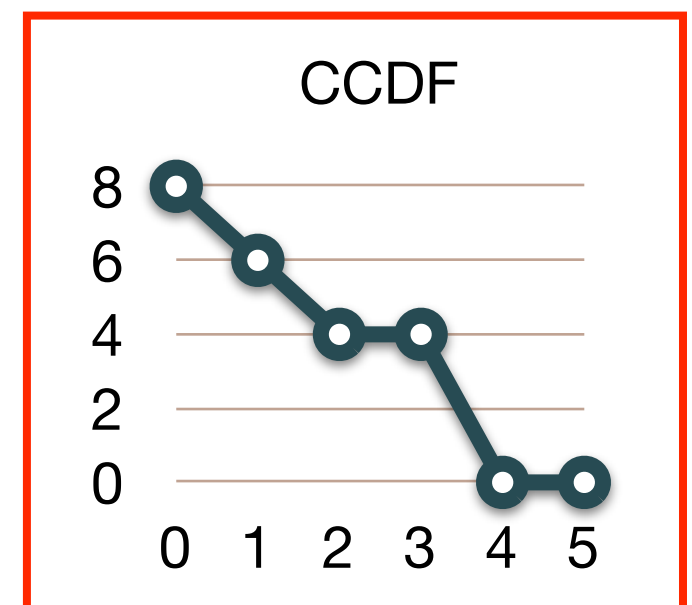
- Degree distribution: the frequency of each degree in graph.
- A widely studied property of networks.

G

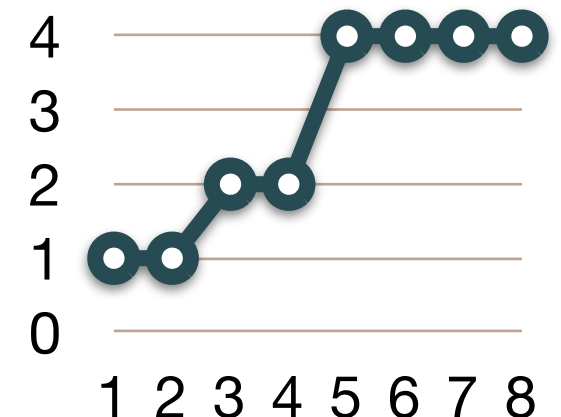


Degree sequence
as a vector

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



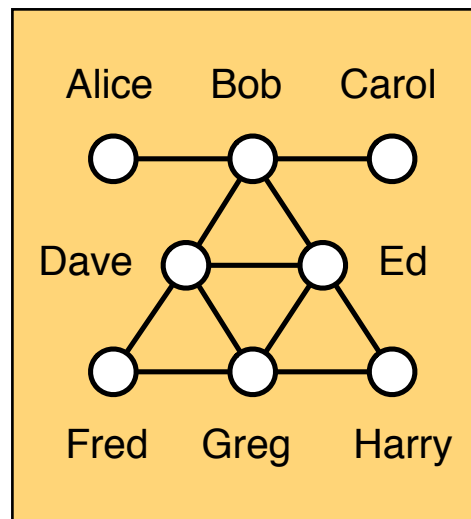
Degree sequence



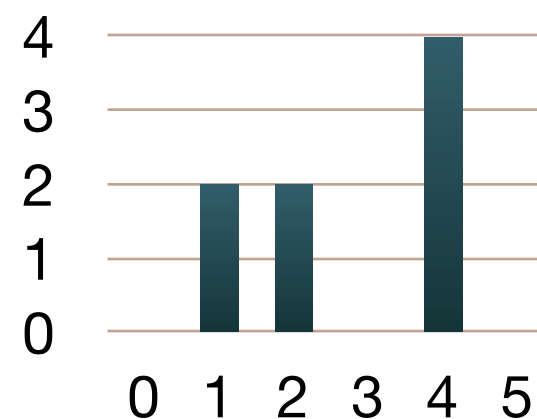
Accurate degree distribution estimation is possible

- Degree distribution: the frequency of each degree in graph.
- A widely studied property of networks.

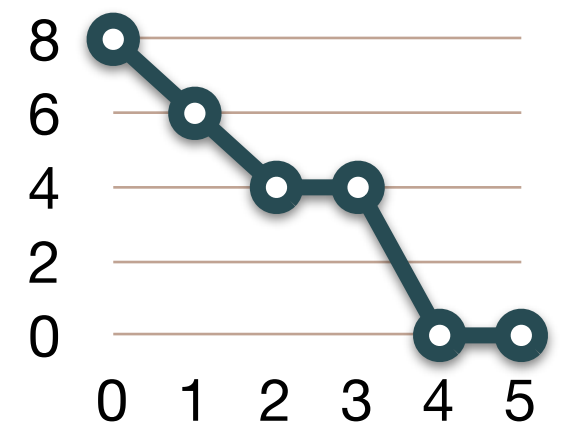
G



Histogram

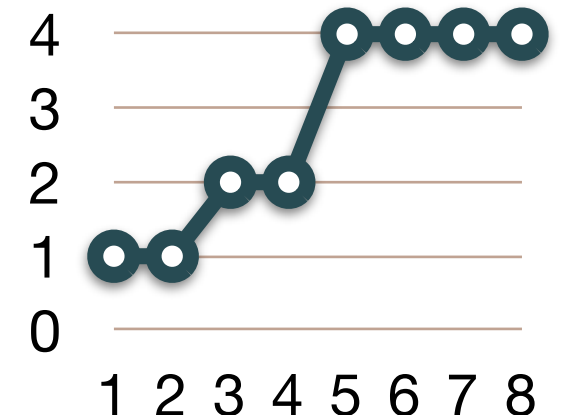


CCDF

Degree sequence
as a vector

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

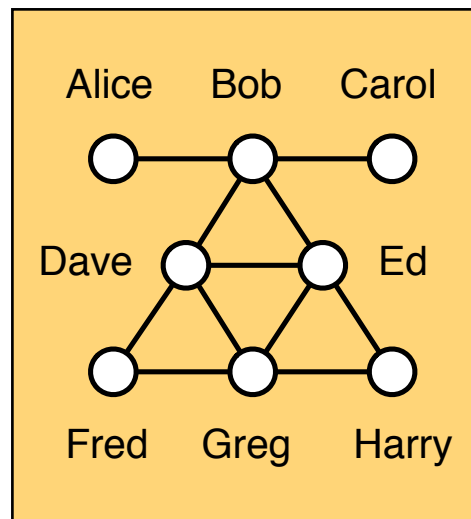
Degree sequence



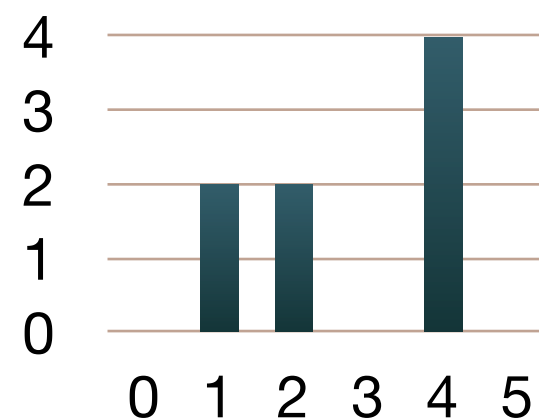
Accurate degree distribution estimation is possible

- Degree distribution: the frequency of each degree in graph.
- A widely studied property of networks.

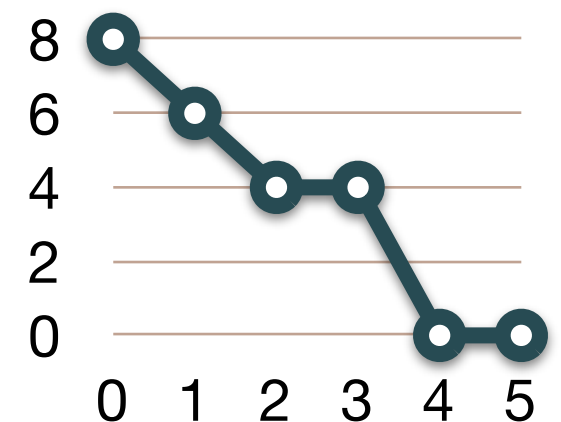
G



Histogram

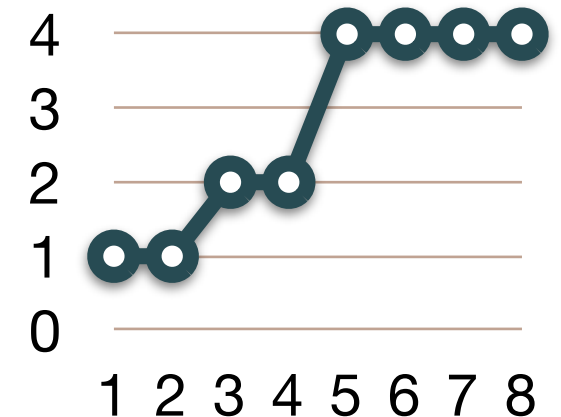


CCDF

Degree sequence
as a vector

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

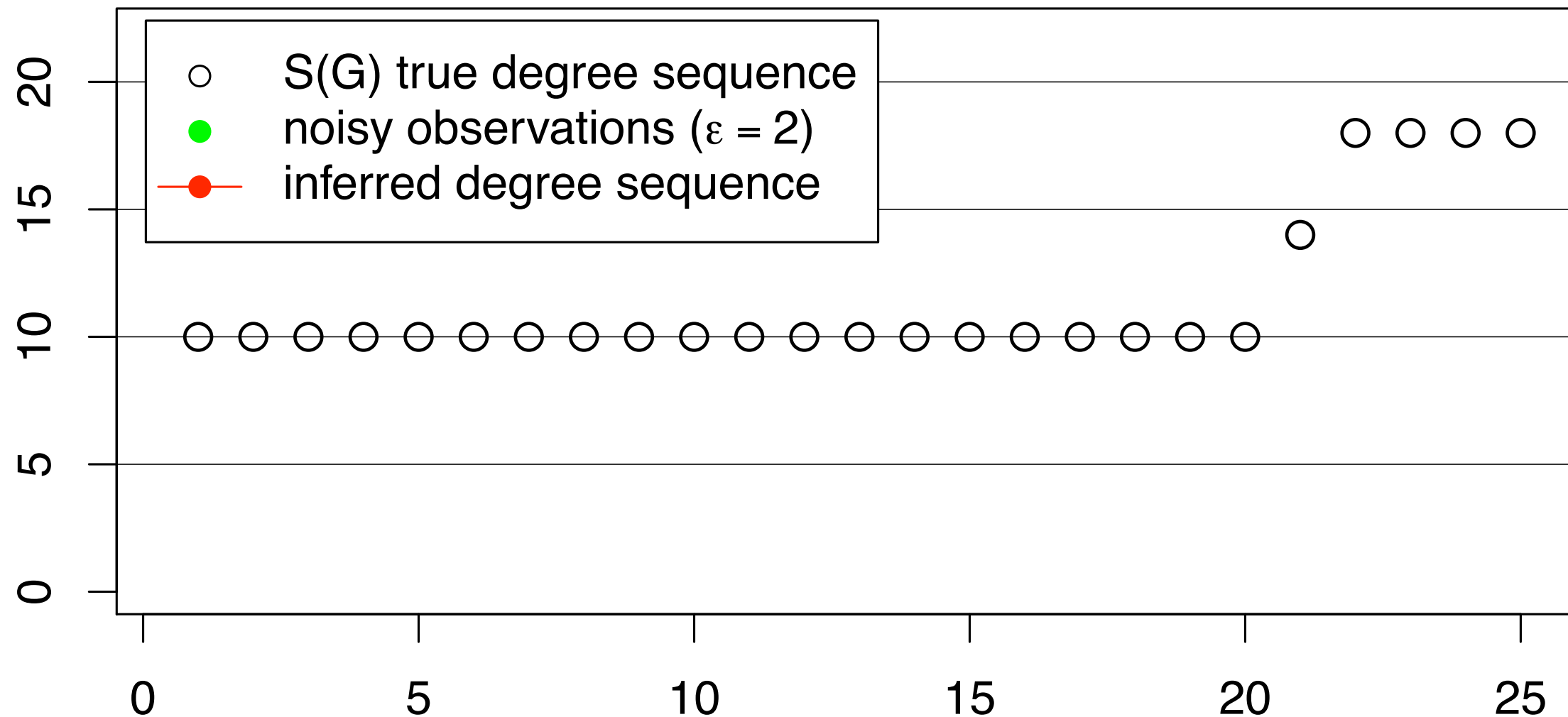
Degree sequence



Using the sort constraint

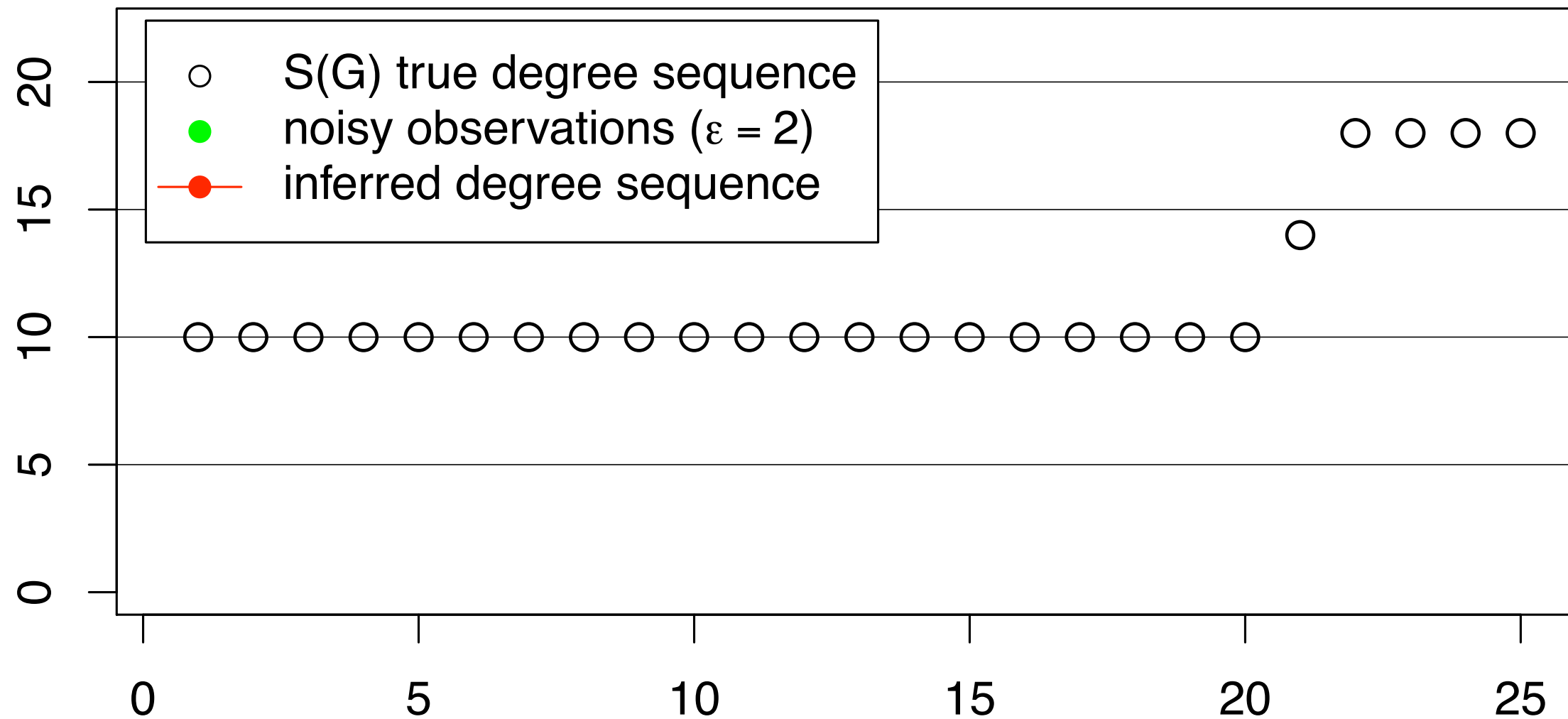
$S(G) = [10, 10, \dots, 10, 10, 14, 18, 18, 18, 18]$

Using the sort constraint

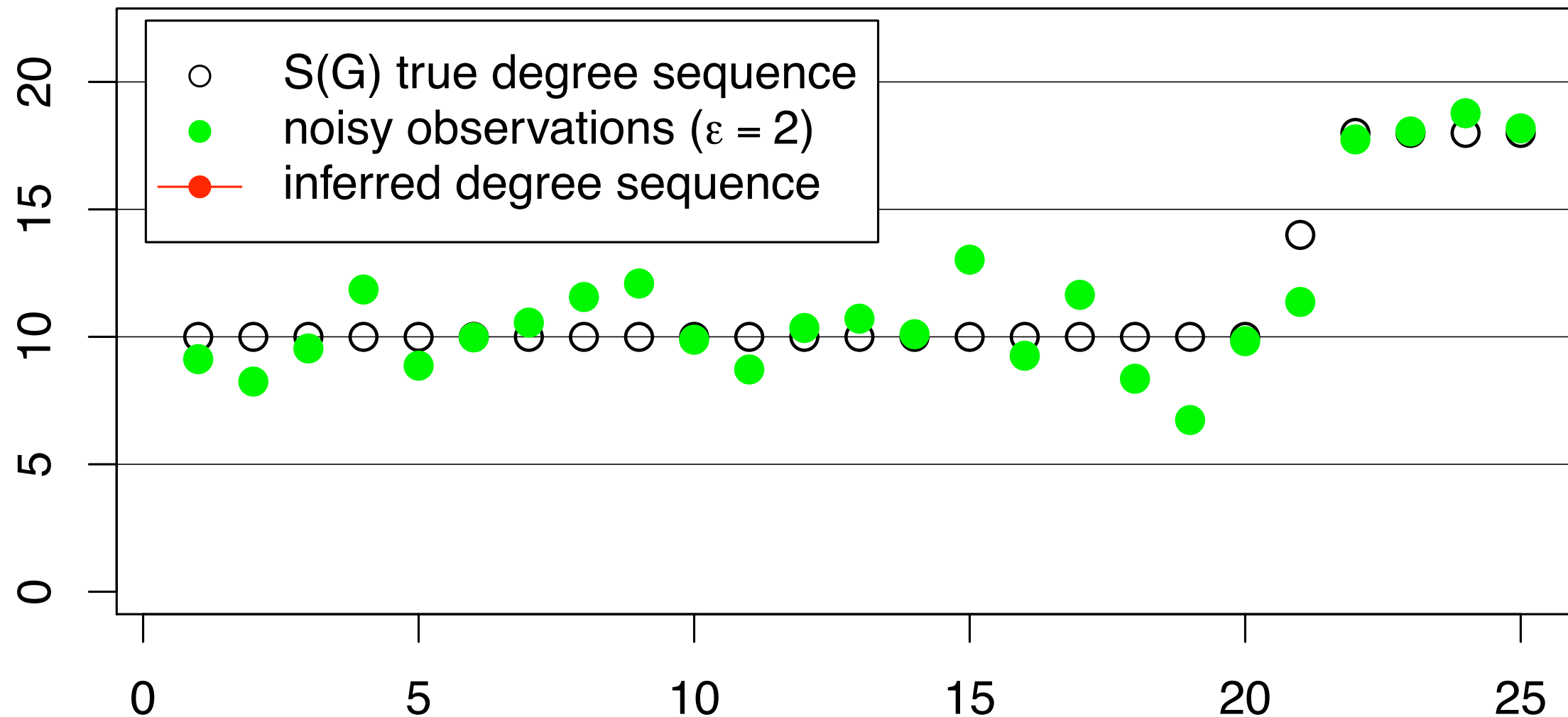


$S(G) = [10, 10, \dots, 10, 10, 14, 18, 18, 18, 18]$

Using the sort constraint

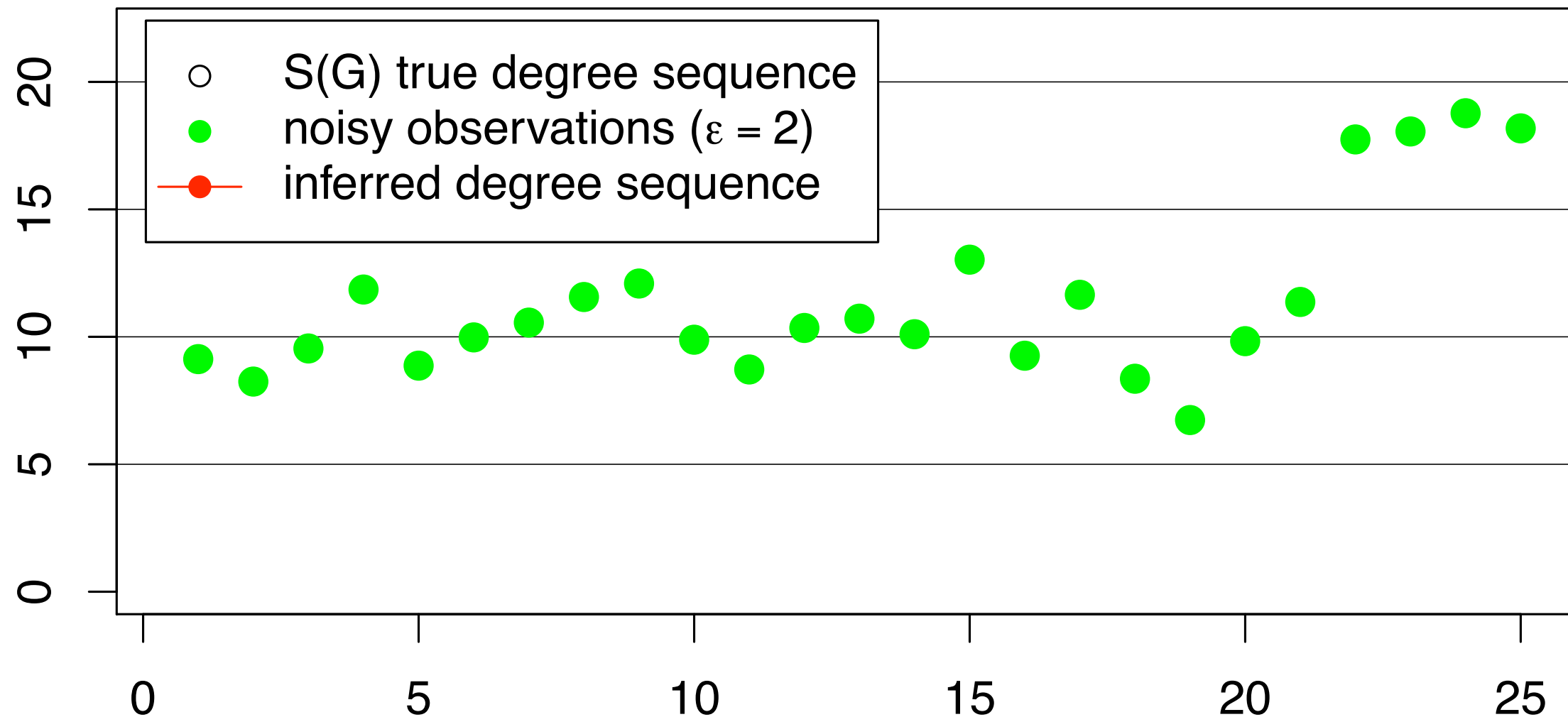


Using the sort constraint



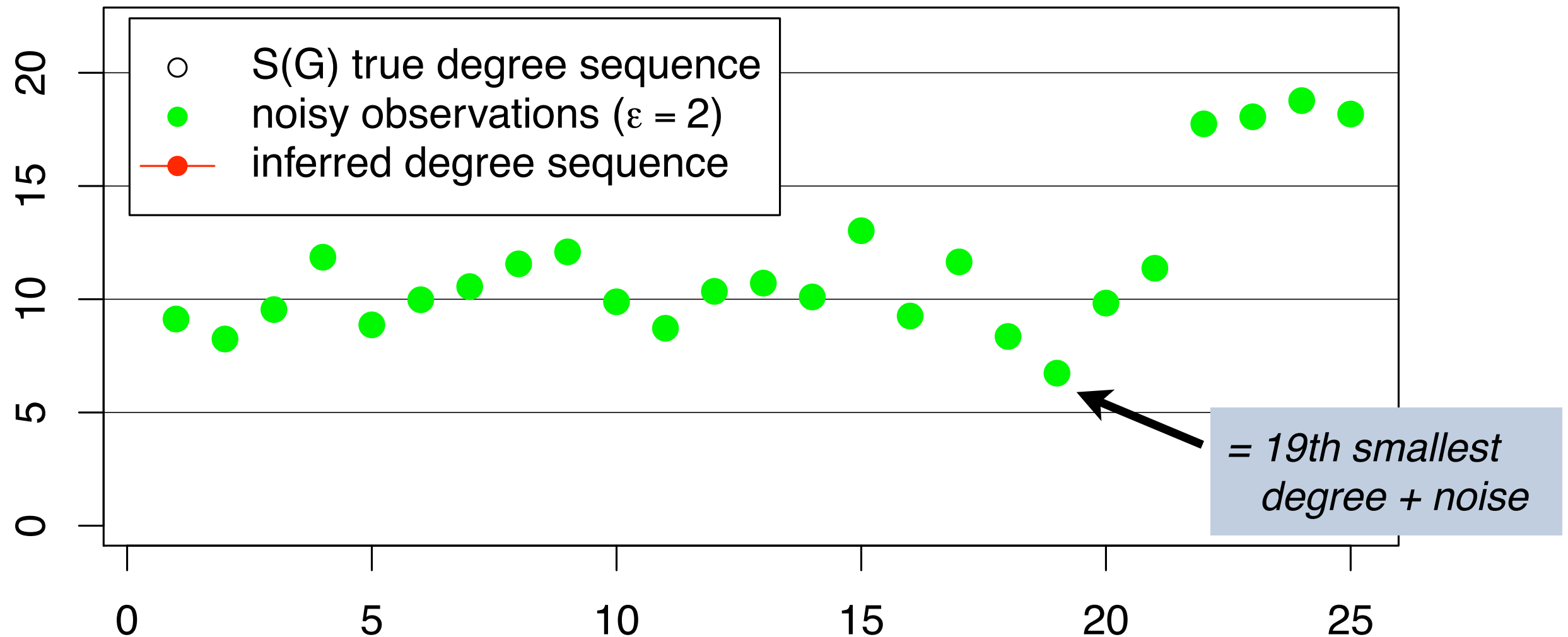
- The output of the sorted degree query is not (in general) sorted.
- We derive a new sequence by computing the **closest** non-decreasing sequence: i.e. minimizing L2 distance.

Using the sort constraint



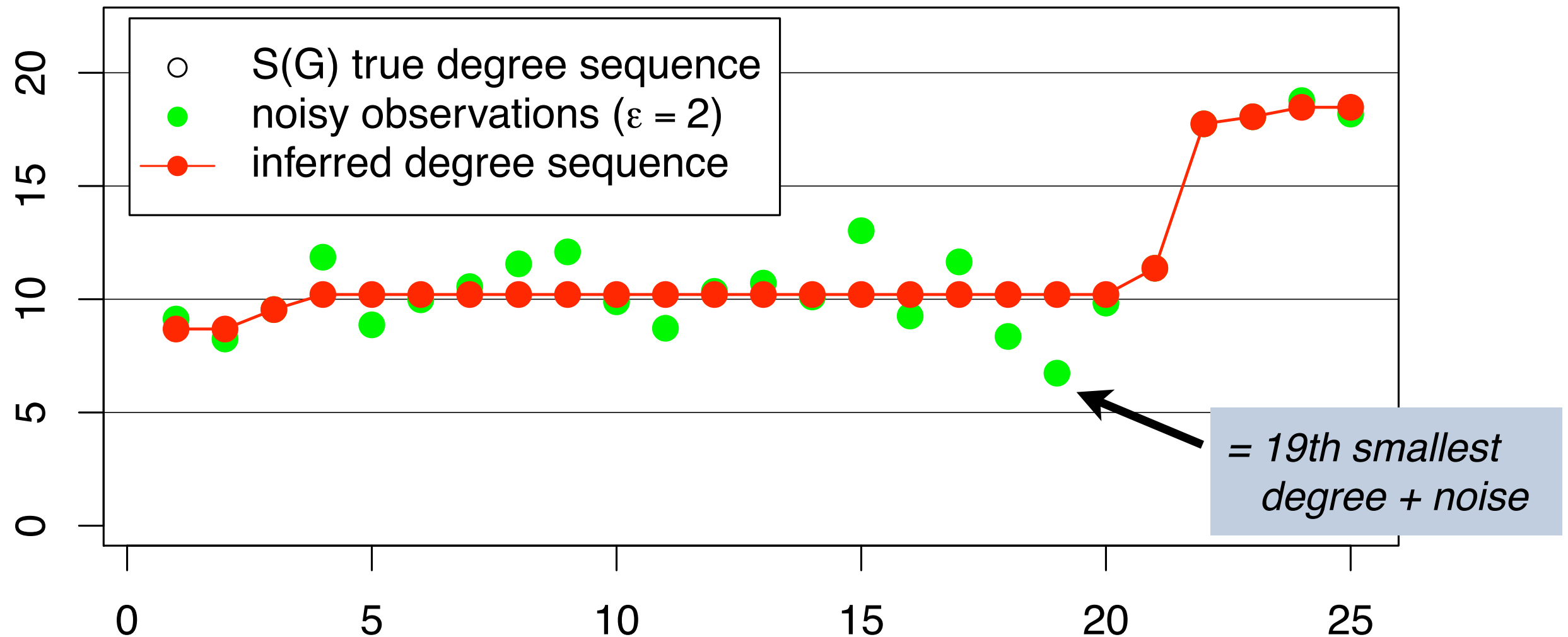
- The output of the sorted degree query is not (in general) sorted.
- We derive a new sequence by computing the **closest** non-decreasing sequence: i.e. minimizing L2 distance.

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.

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

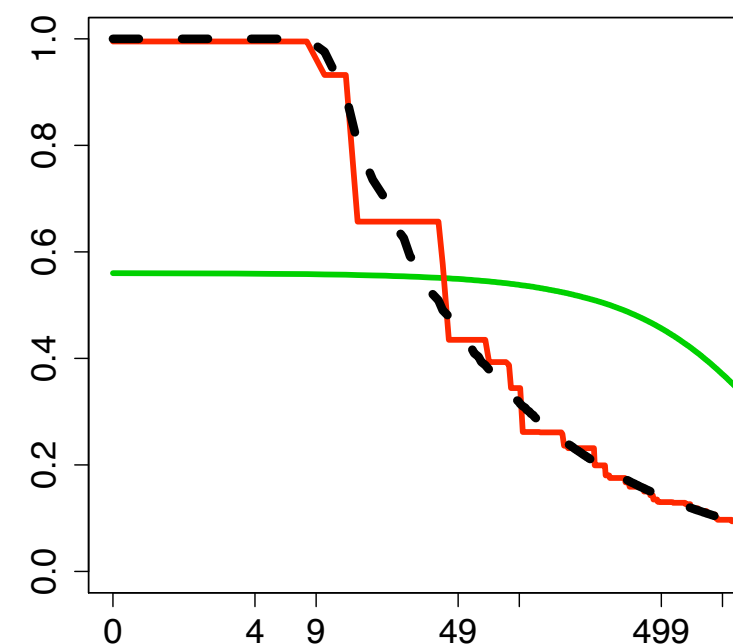
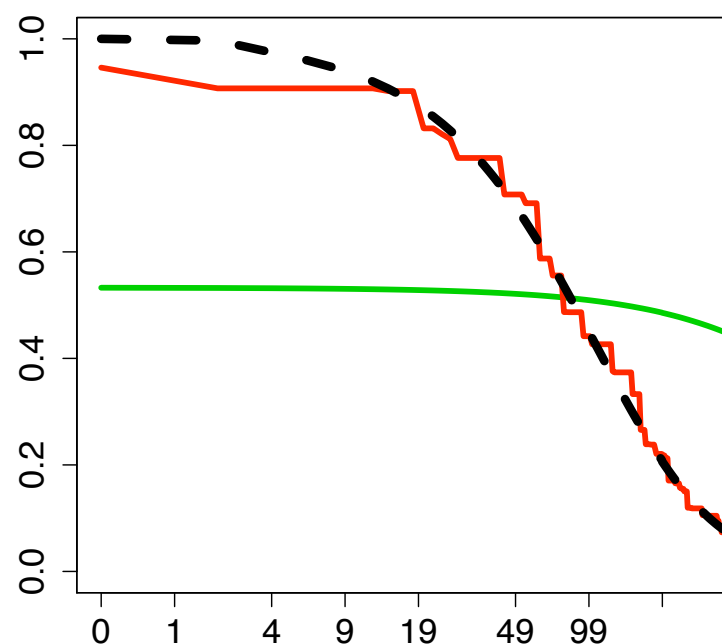
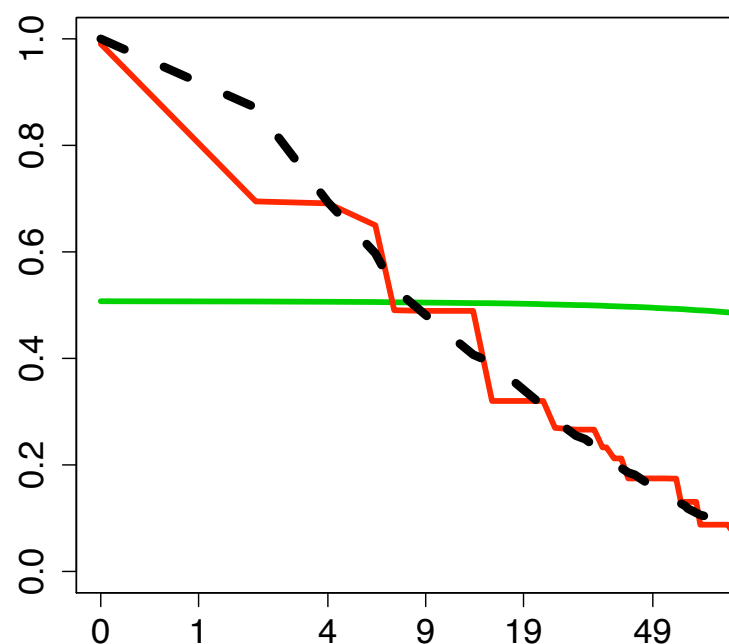
original 
noisy 
inferred 

livejournal
n=5.3M

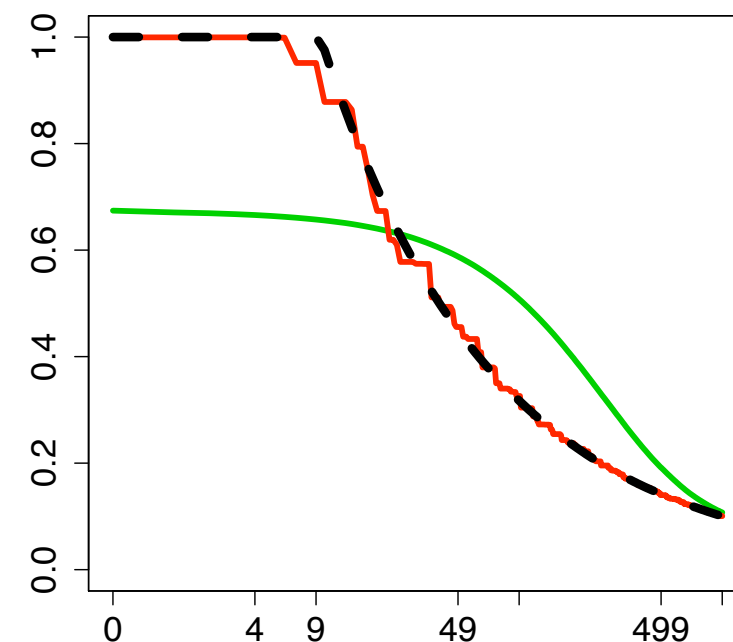
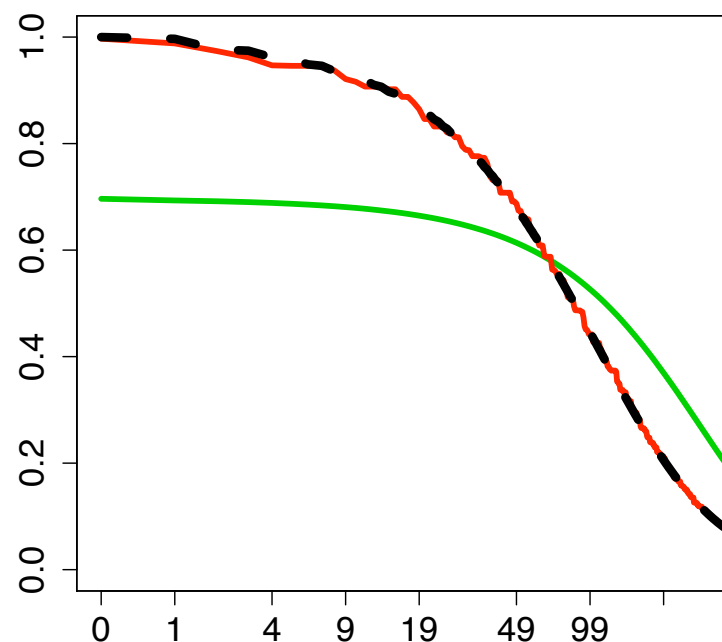
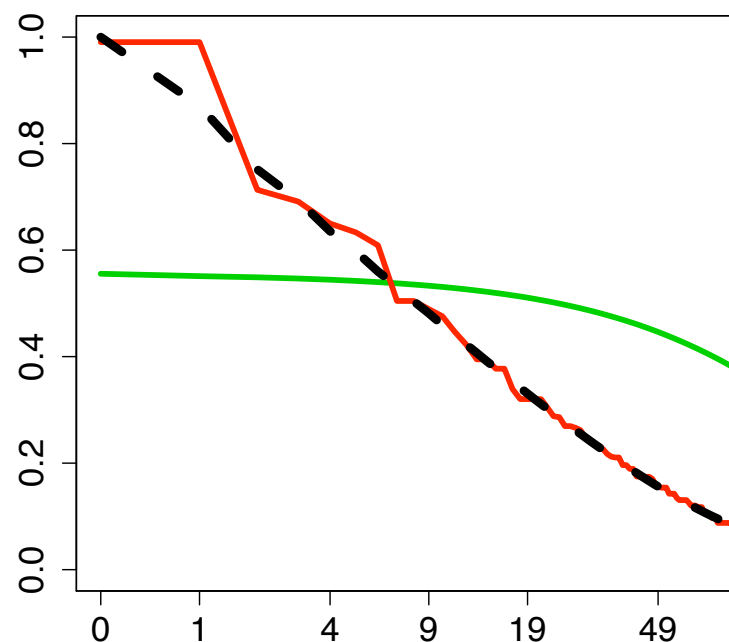
orkut
n=3.1M

powerlaw
 $\alpha=1.5$, n=5M

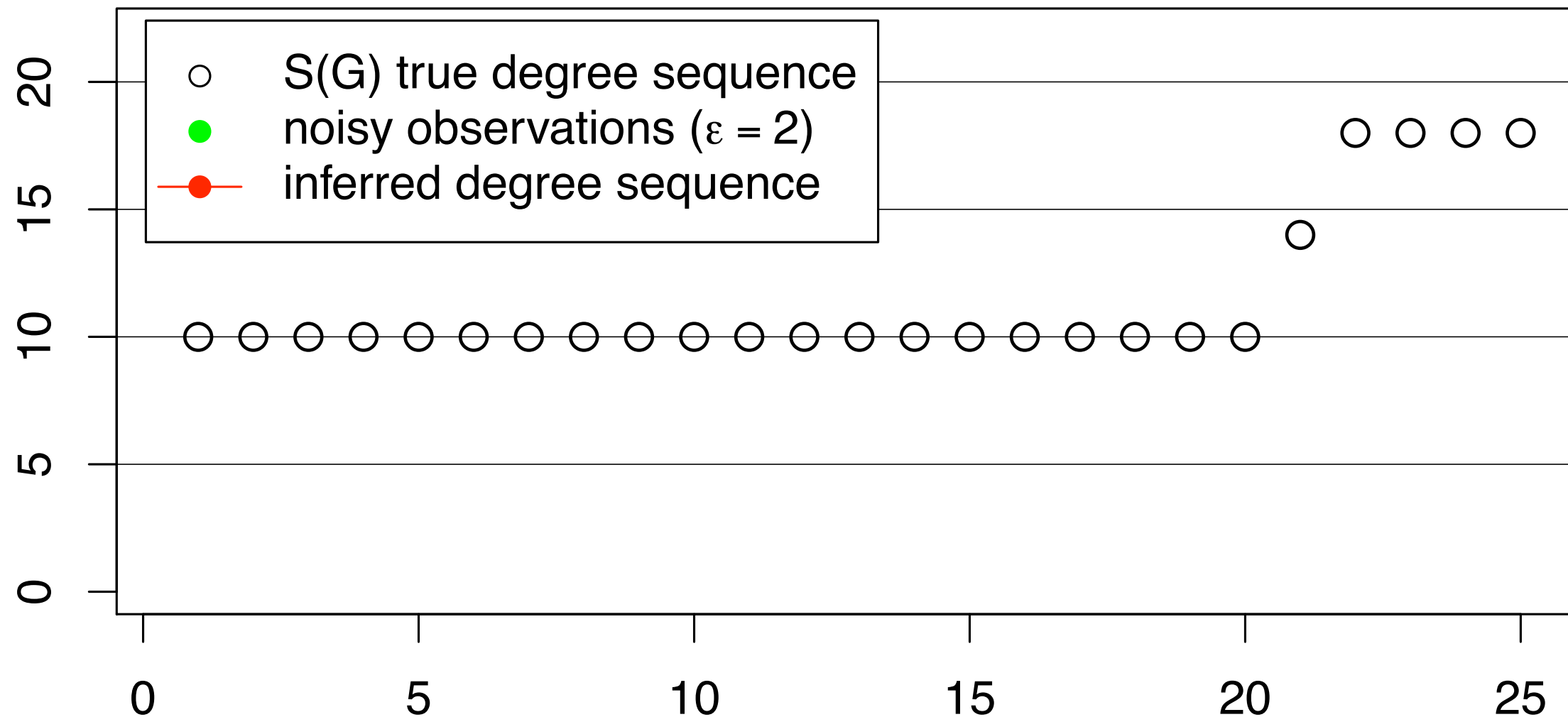
$\epsilon=.001$



$\epsilon=.01$

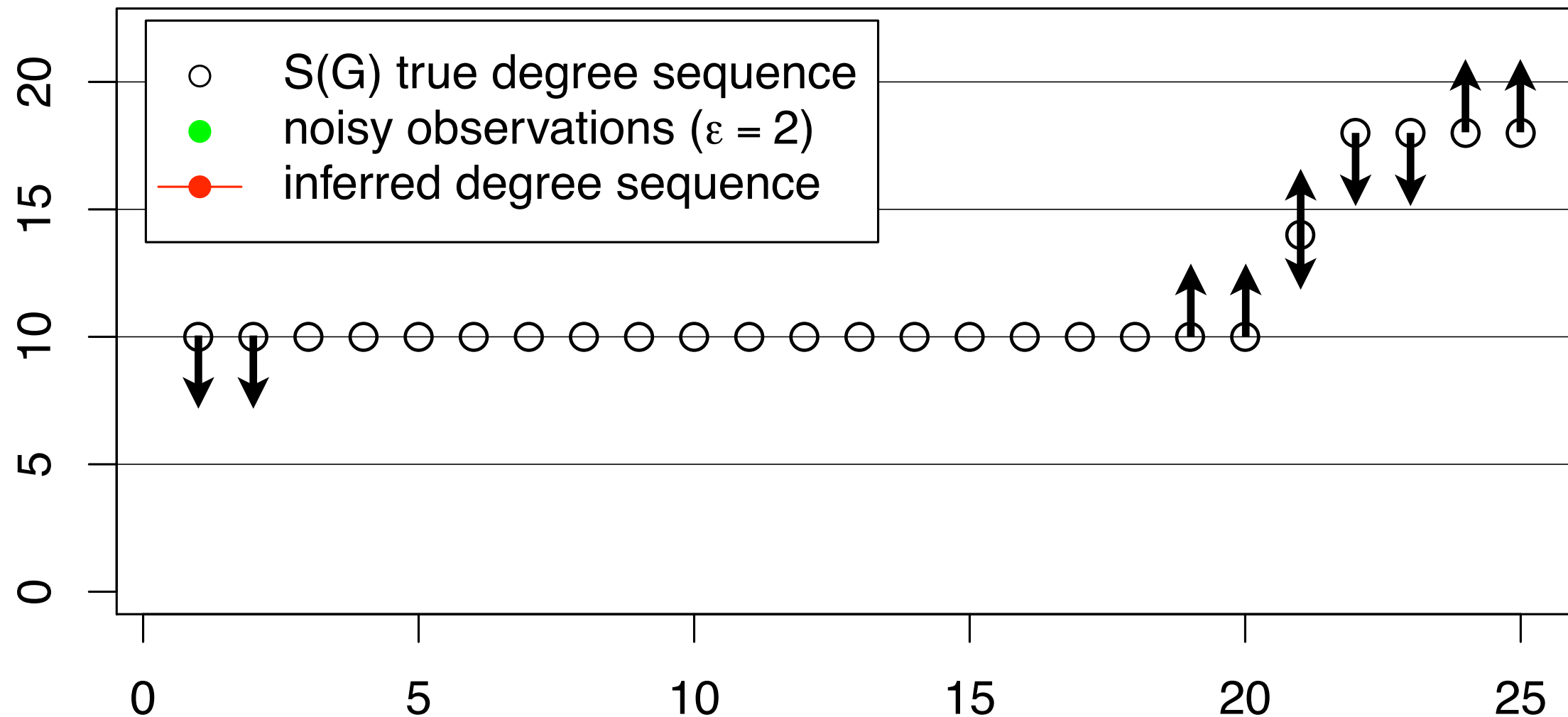


After inference, noise only where needed



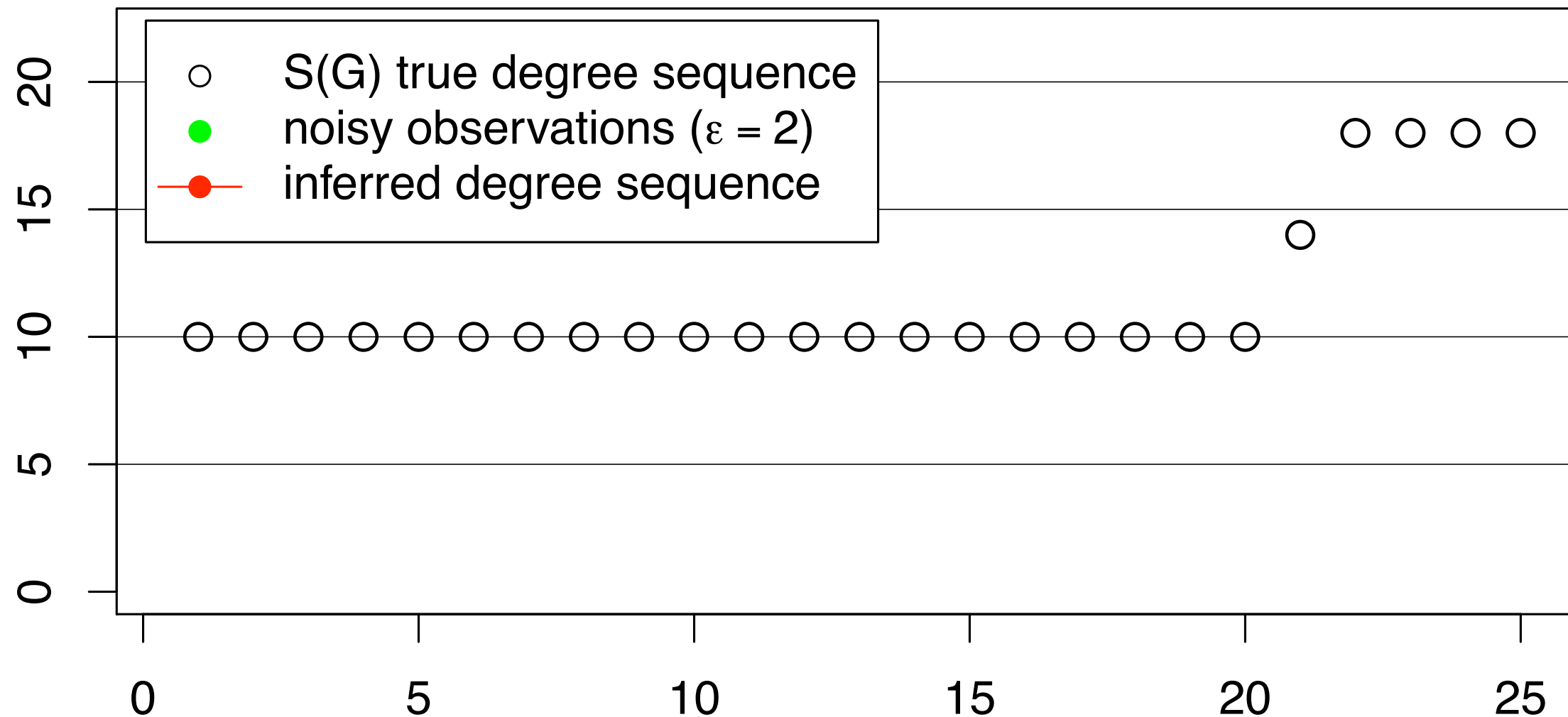
- Standard Laplace noise is sufficient *but not necessary* for differential privacy.
- By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.
 - Improvement in accuracy will depend on sequence

After inference, noise only where needed



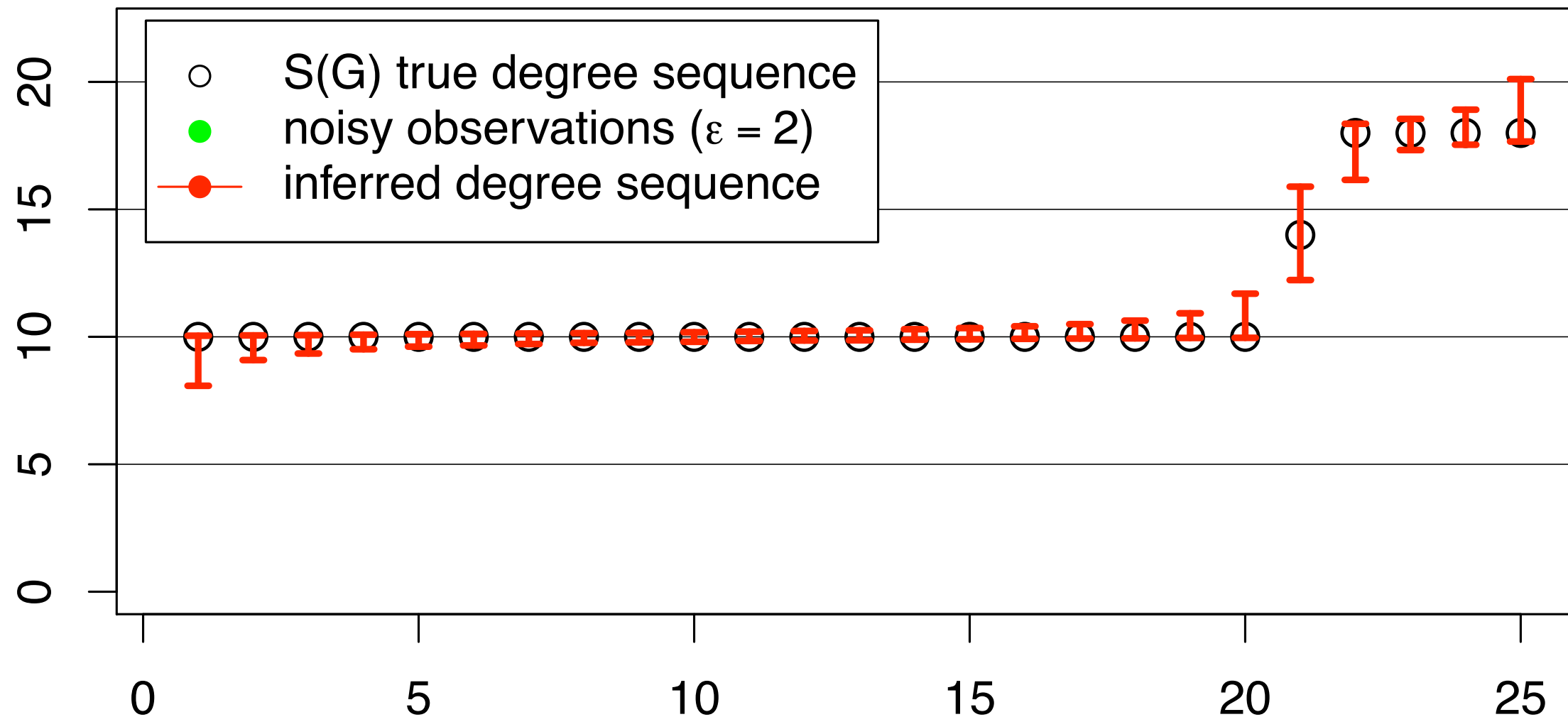
- Standard Laplace noise is sufficient *but not necessary* for differential privacy.
- By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.
 - Improvement in accuracy will depend on sequence

After inference, noise only where needed



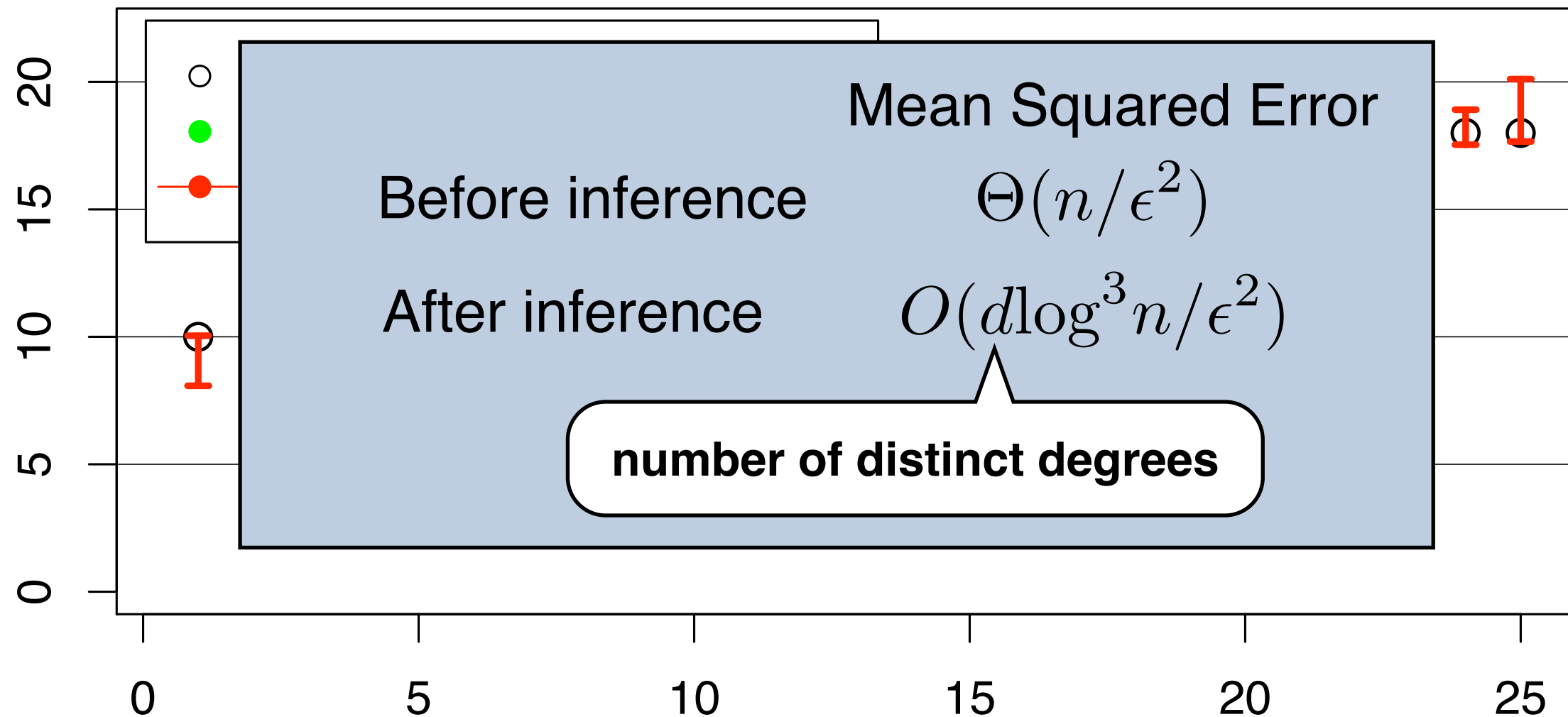
- Standard Laplace noise is sufficient *but not necessary* for differential privacy.
- By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.
 - Improvement in accuracy will depend on sequence

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

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

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:

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

<http://www.cs.umass.edu/~mhay/>

References

- **[Blum, PODS 05]** A. Blum C. Dwork, F. McSherry, and K. Nissim. Practical Privacy: The SuLQ Framework. PODS, 2005.
- **[Blum, STOC 08]** A. Blum, K. Ligett, and A. Roth. Learning Theory Approach to Non-Interactive Database Privacy, STOC 2008.
- **[Onnela, PNAS 07]** J. Onnela et al. Structure and tie strengths in mobile communication networks, PNAS 2007.
- **[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.

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.