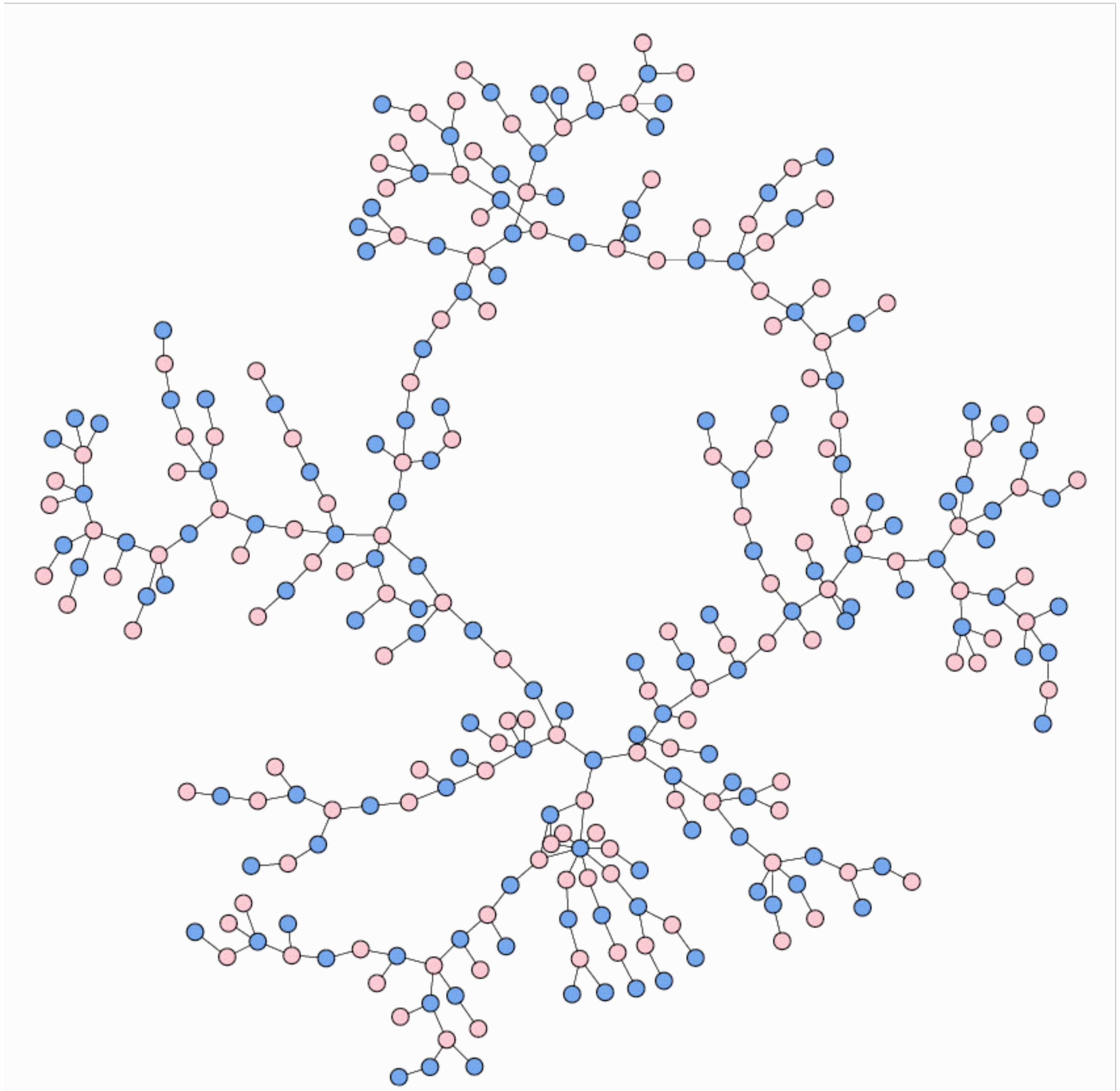
An Introduction to the Private Analysis of Network Data

Michael Hay, Colgate University
Gerome Miklau, University of Massachusetts, Amherst

Charles River Workshop on Private Analysis of Social Networks, May 2014.

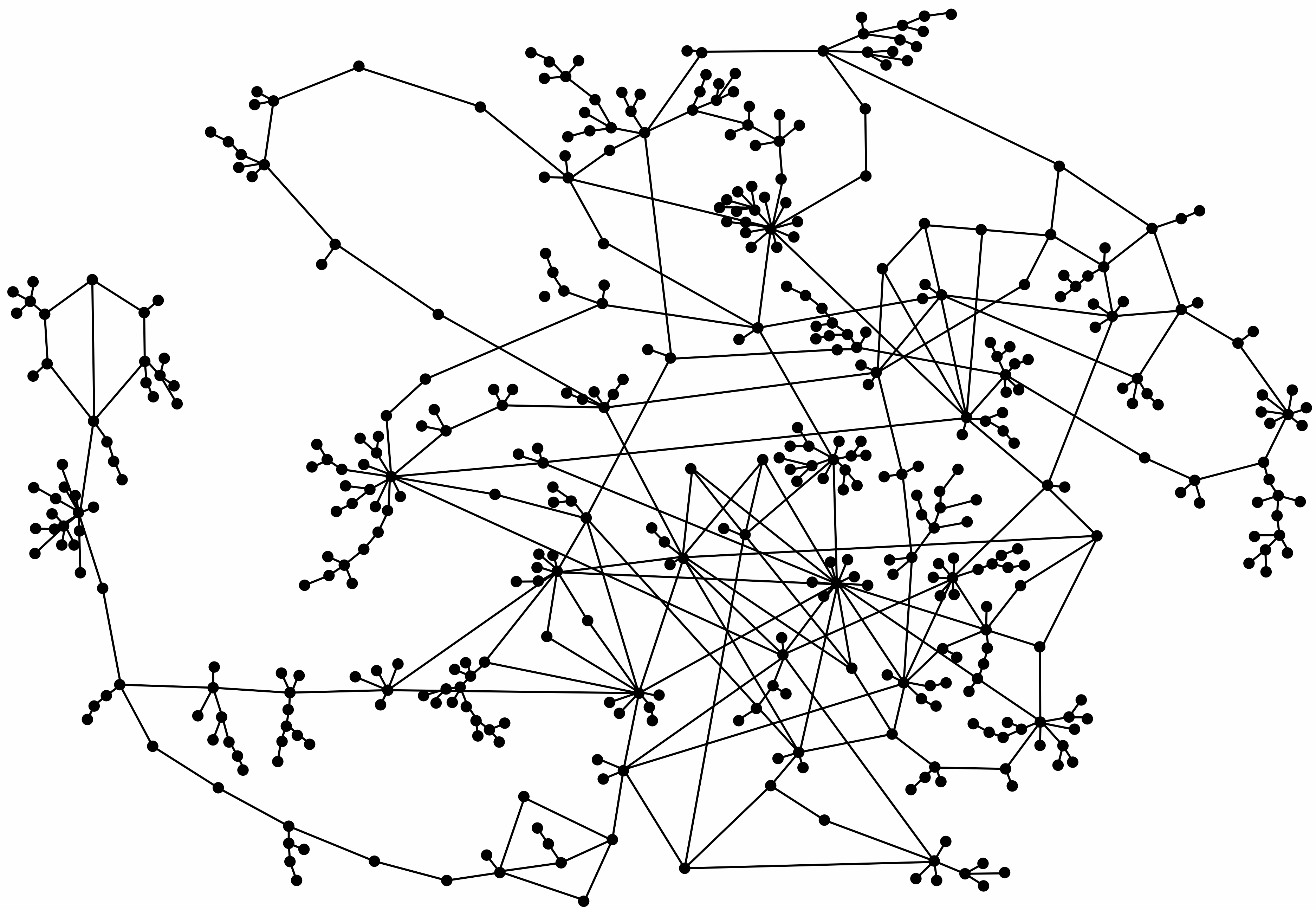
Romantic connections in a high school



Bearman, et al.
The structure of adolescent romantic and sexual networks.
American Journal of Sociology, 2004.

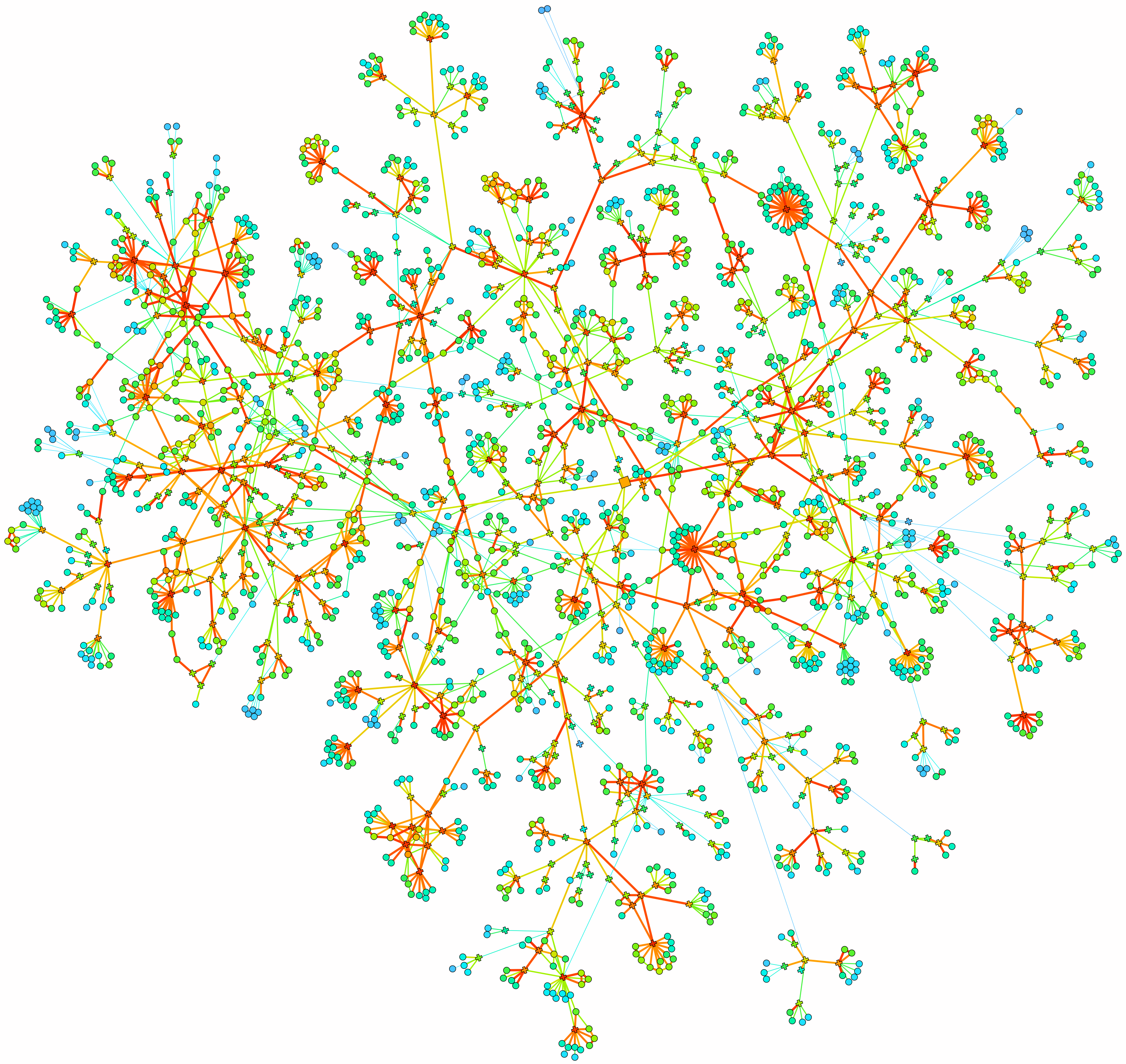
(Image drawn by Newman)

Sexual and injecting drug partners



Potterat, et al.
Risk network structure in the early epidemic phase of hiv transmission in colorado springs.
Sexually Transmitted Infections, 2002.

Social ties derived from a mobile phone network



J. Onnela et al.
Structure and tie strengths in mobile communication networks,
Proceedings of the National Academy of Sciences, *2007*

Sensitive data

Information about an individual that deserves protection because its release could cause harm.

A tabular data model

identifier

descriptive (sensitive) attributes

ID	Age	HIV
Alice	25	Pos
Bob	19	Neg
Carol	34	Pos
Dave	45	Pos
Ed	32	Neg
Fred	28	Neg
Greg	54	Pos
Harry	49	Neg

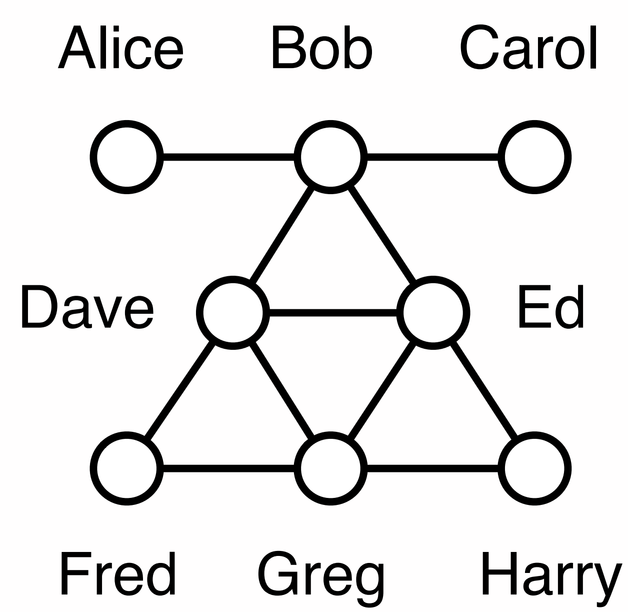
independence:
Bob's record doesn't reveal anything about Fred

Sensitive fact: "Greg's HIV status is positive"

A network data model

Nodes

ID	Age	HIV
Alice	25	Pos
Bob	19	Neg
Carol	34	Pos
Dave	45	Pos
Ed	32	Neg
Fred	28	Neg
Greg	54	Pos
Harry	49	Neg



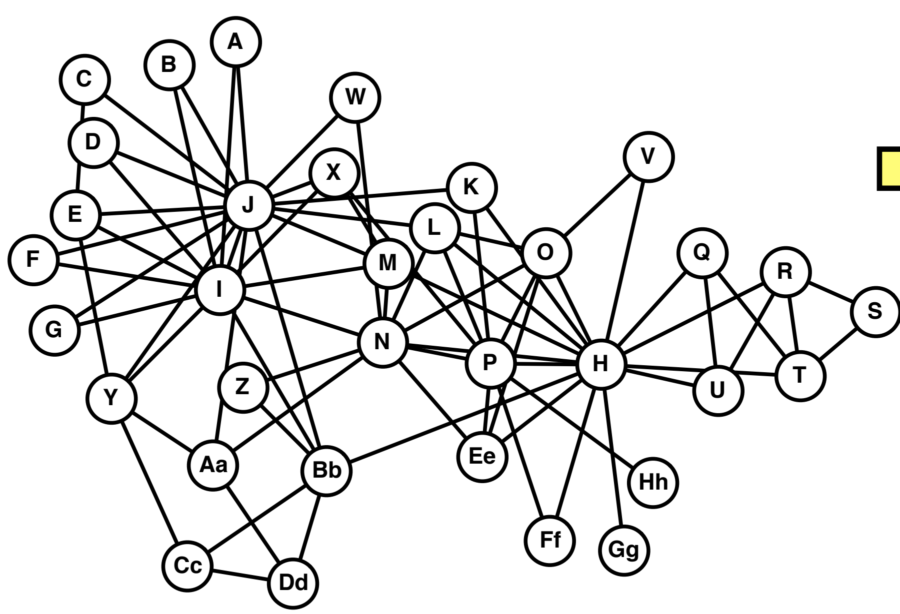
Edges

ID1	ID2
Alice	Bob
Bob	Carol
Bob	Dave
Bob	Ed
Dave	Ed
Dave	Fred
Dave	Greg
Ed	Greg
Ed	Harry
Fred	Greg
Greg	Harry

Sensitive facts:
“Greg is connected to Ed.”
“Greg is connected to 4 people.”
“Greg is connected to one HIV positive person.”
“Greg’s friends tend to be connected to one another.”
….

Problem setting

DATA OWNER	ANALYST / ADVERSARY
(trusted)	(untrusted)



sensitive data set

✔

"global" properties

"How rapidly do rumors spread in this network?"

"Are people most likely to form friendships with those who share their attributes?"

✖

sensitive facts

Can we enable analysts to study useful properties of networks without revealing sensitive facts?

Approaches that don’t work (or don’t work well)

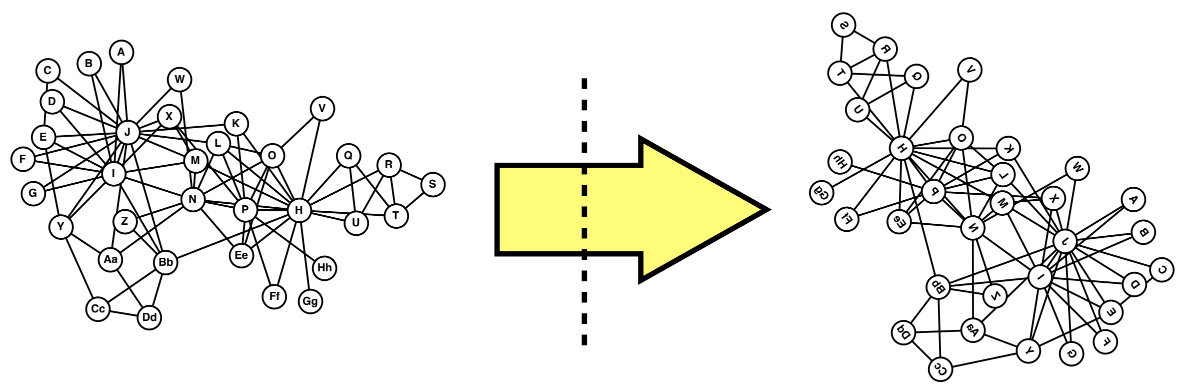
- **Access control**: grant/revoke access to data objects
- **Releasing “aggregate” information.**
- **Query auditing**: start answering queries (truthfully), but stop when they become dangerous.
- **Sampling**: include only a fraction of respondents’ data
- **Anonymization/Sanitization:** remove identifiers from respondent’s data

Private analysis of social networks

- Competing goals
  - “Utility”: analysts can measure global properties accurately
  - “Privacy”: sensitive facts not disclosed
- Typical problem formulation in privacy research:
  - Formally define privacy condition: “safe for release”
  - **Guarantee privacy**: provable privacy condition (worst-case assumptions)
  - **Measure utility**: establish error bounds, empirical studies (average case)

Methods of release

- **Data publishing**

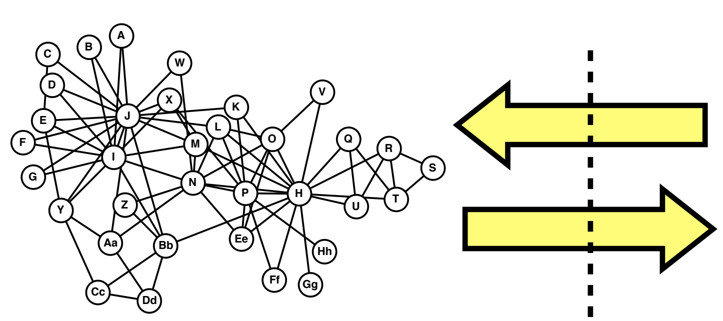


sensitive data set **safe data set**

Data transformed to make safe to release

- appealing to analyst
- utility more limited than it may appear.

- **Query answering**



queries

safe answers

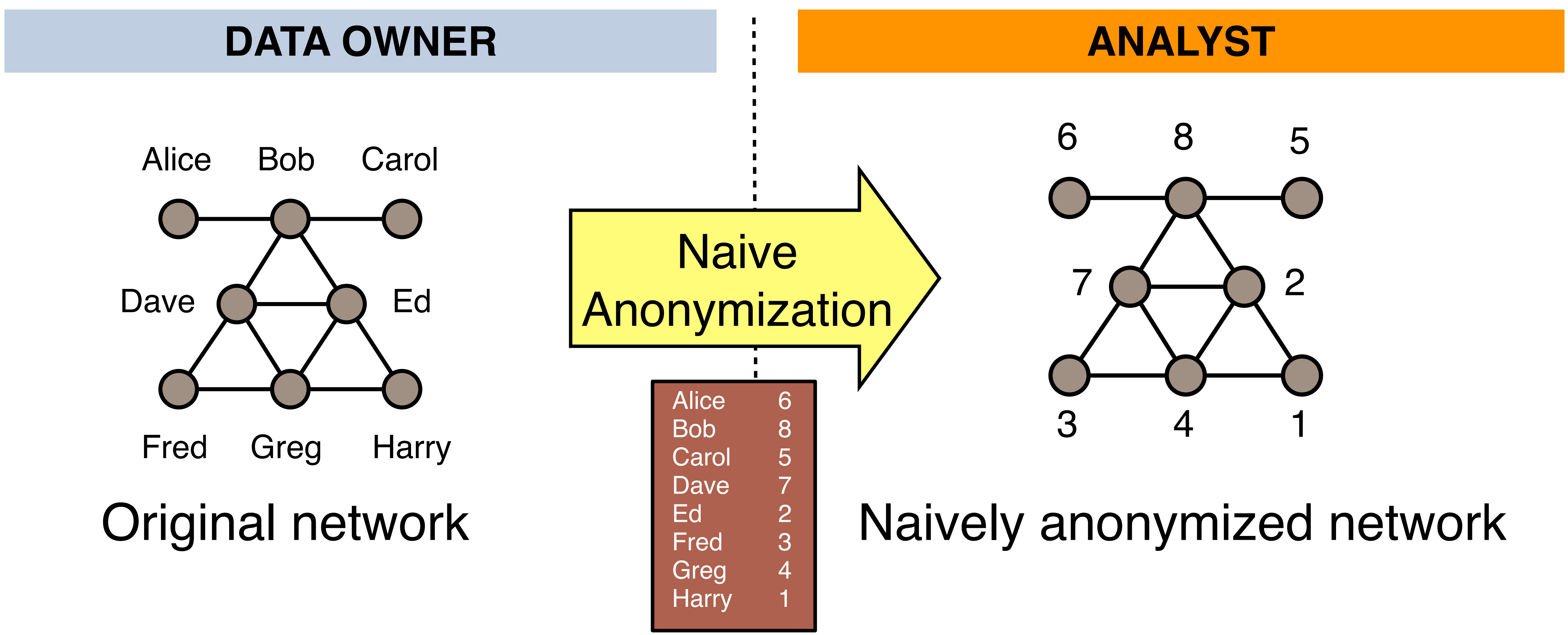
sensitive data set

Answers altered to make safe (e.g., random noise added)

- analyst's interaction with data is limited
- good solutions for specific classes of queries

Naive anonymization

Naive anonymization is a transformation of the network in which identifiers are replaced with random numbers.



Good utility: output is isomorphic to the original network

Adversaries with **external information**

External information: facts about *identified* individuals and their relationships in the hidden network.

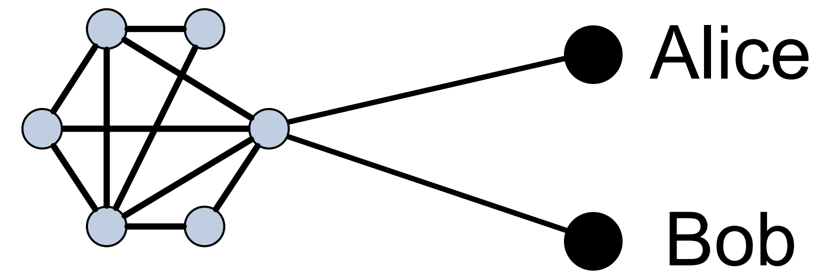
- Sources of external information
  - Background information (web, public records, etc.)
  - Related publicly-available data sets (auxiliary network attack)
  - Adversary may be network participant!
- Brief but colorful history of attacks on *real* anonymized data
  - Medical records **[Sweeney 00]**, search engine logs **[Barbaro 96]**, netflix movie ratings **[Narayan 06]**, genetic data **[Homer 08]**, …
- Illustrative example: active attack on network data

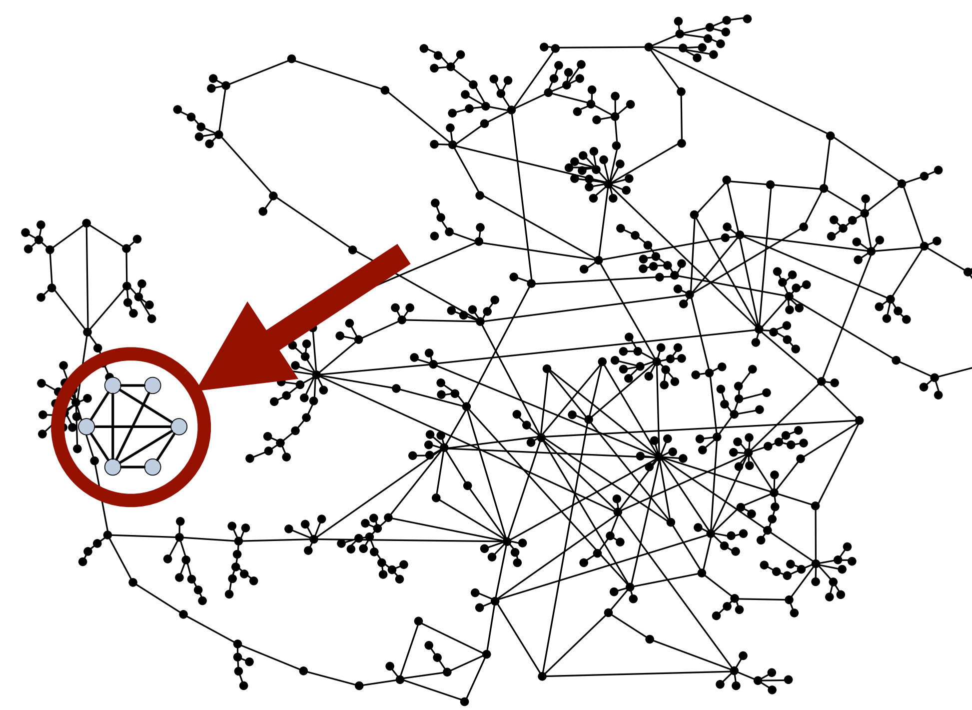
Active attack

- Goal: **disclose edge** between two targeted individuals.
- Key assumption: adversary can alter the network structure, by creating nodes and edges, **prior to** naive anonymization.
  - In blogging network: create new blogs and links to other blogs.
  - In email network: create new identities, send mail to identities.
  - (Harder to carry out this attack in a social network where “friendship” connection must be reciprocated by target.)

Active attack on an online network

1	Attacker creates a distinctive **subgraph** of nodes and edges.
2	Attacker links subgraph to target nodes in the network.
Naive anonymization	
3	Attacker finds matches for pattern in naively anonymized network.
4	Attacker re-identifies targets and discloses structural properties.





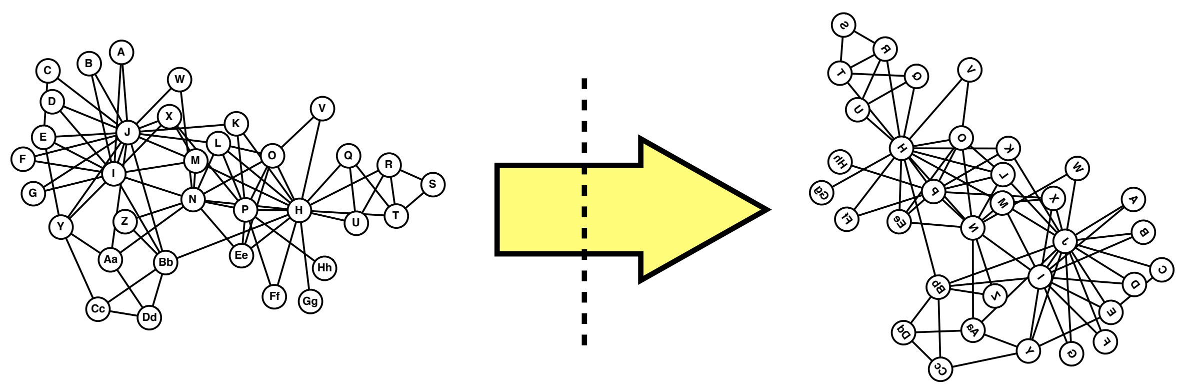
Results
- Subgraph can be small (inconspicuous)
- Does not require knowledge of input graph
- Attack likely to succeed w.h.p.

Response to failure of anonymization

- Given limitations of naive anonymization, much work on more aggressive forms of anonymization **[survey: Hay, Privacy-Aware Knowledge Discovery 10]**
  - Network structure altered to prevent certain attacks
  - Safety criteria is defined in terms of resistance to (known) attacks.
- Looming concern: vulnerability to unanticipated attacks.
- History (for tabular data anonymization) of published techniques later shown to be vulnerable to attack **[survey: Chen, Foundations and Trends in Database 09]**
- We need more **rigorous safety criteria**

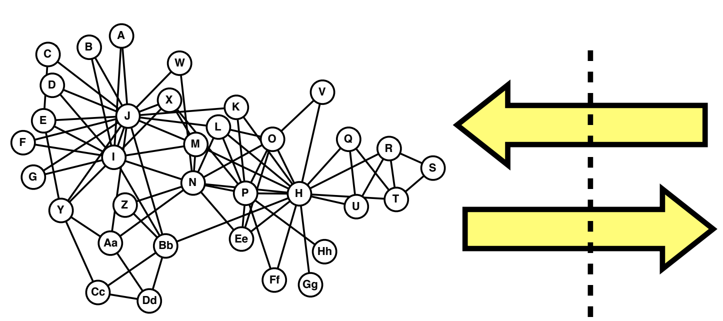
Methods of release

- **Data publishing**



sensitive data set **safe data set**

- **Query answering**



queries

safe answers

sensitive data set

Queries typically aggregate network statistics. Examples:

- degree distribution
- subgraph counts

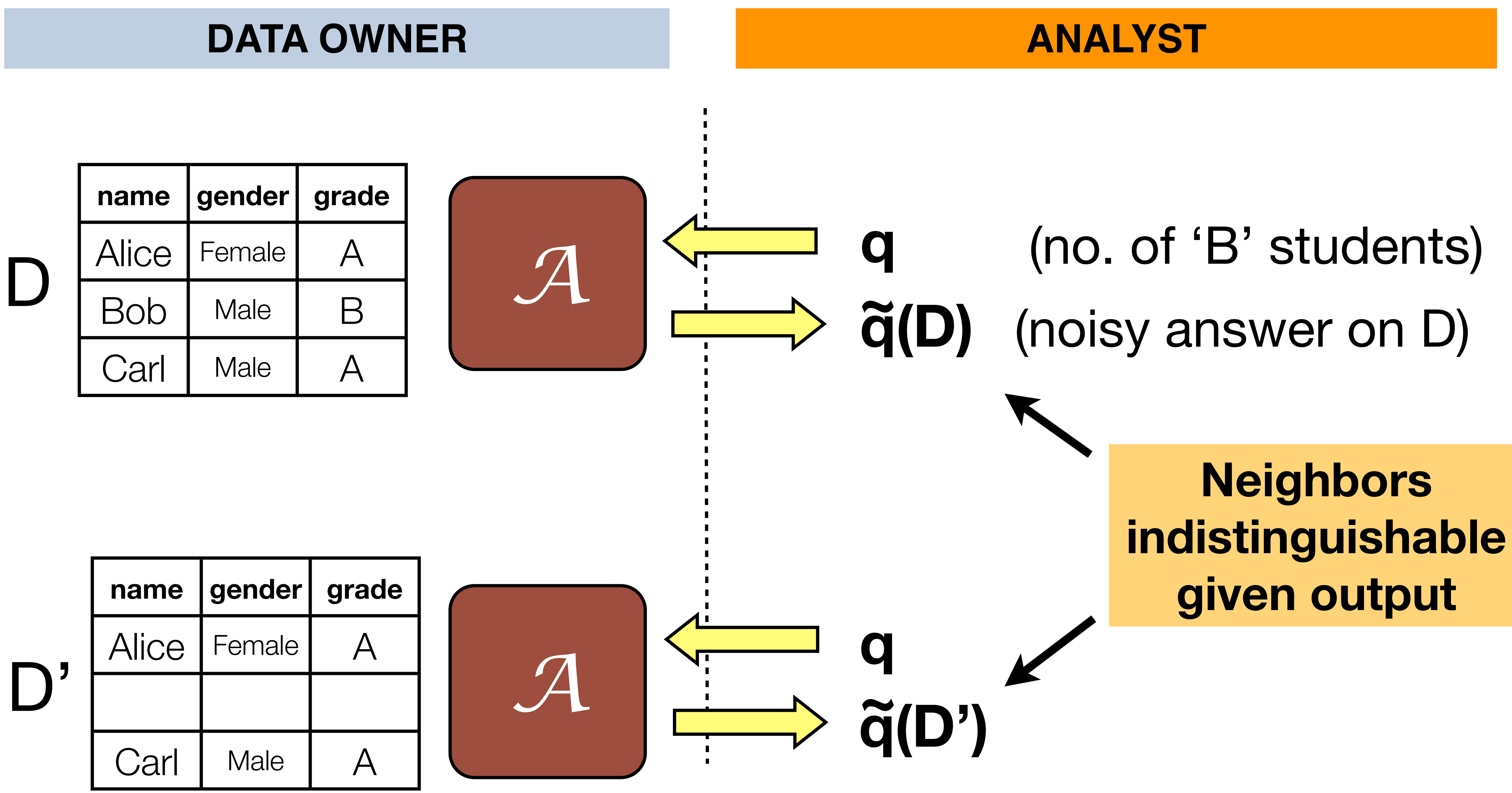
The differential guarantee

D

name	gender	grade
Alice	Female	A
Bob	Male	B
Carl	Male	A

D’

name	gender	grade
Alice	Female	A
Carl	Male	A



Two databases are **neighbors** if they differ by at most one tuple

Query sensitivity

The sensitivity of a query q is

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

where D, D' are any two neighboring databases

q1	Count('B' students)	$\Delta q1 = 1$
q2	Max(Salary of all emps)	$\Delta q2 = (max\text{-}min)$
q3	Count(emps with salary in [450k,500k])	$\Delta q3 = 1$

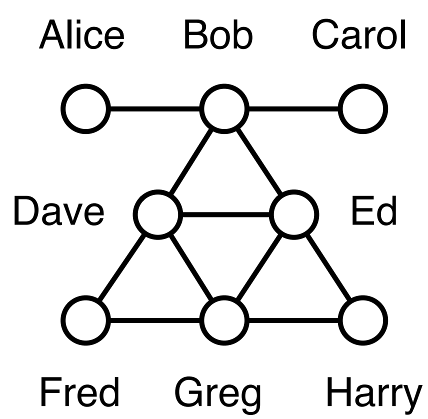
Query sensitivity on network data

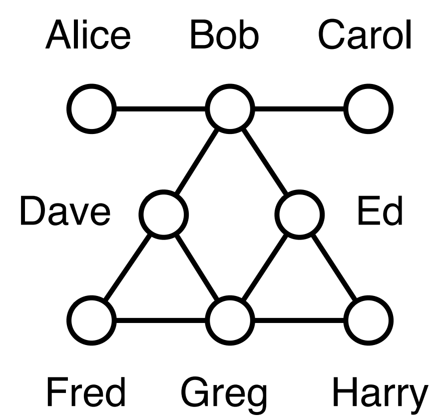
- For tabular data, neighboring databases differ by one record
  - Intuitive rationale: measure how much one person’s data can affect result
- For network data, should neighboring database differ...
  - ... by one record? **(edge sensitivity)**
  - ... by contribution of one person’s data? **(node sensitivity)**
- Choice impacts both privacy and utility

Degree queries have low (edge) sensitivity

- **$Q_{DEGREE=d}$**: return the number of nodes of degree d in the network

G





G'

$Q_{DEGREE=4}$ (G) = 4 **$Q_{DEGREE=4}$ (G') = 2**

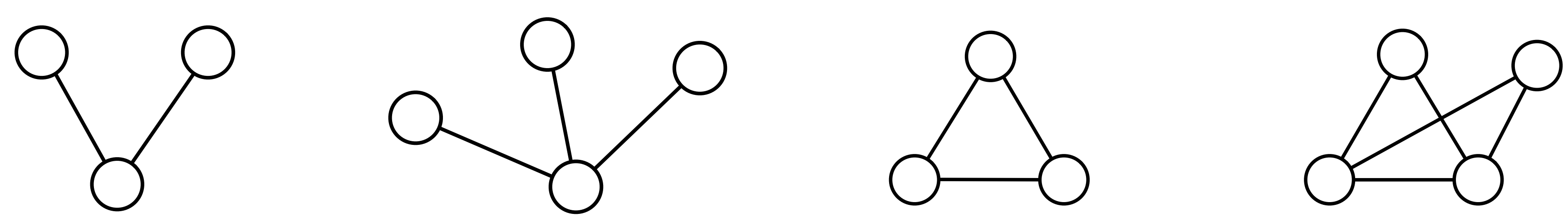
Low Sensitivity:

Δ**Q_{DEGREE}=2**

- Degree distributions (**$Q_{DEGREE=d}$** for all d) can be answered accurately under (edge) differential privacy **[Hay, PVLDB 10]**

Subgraph counting queries

- Given query graph H, return the number of subgraphs of G that are isomorphic to H.

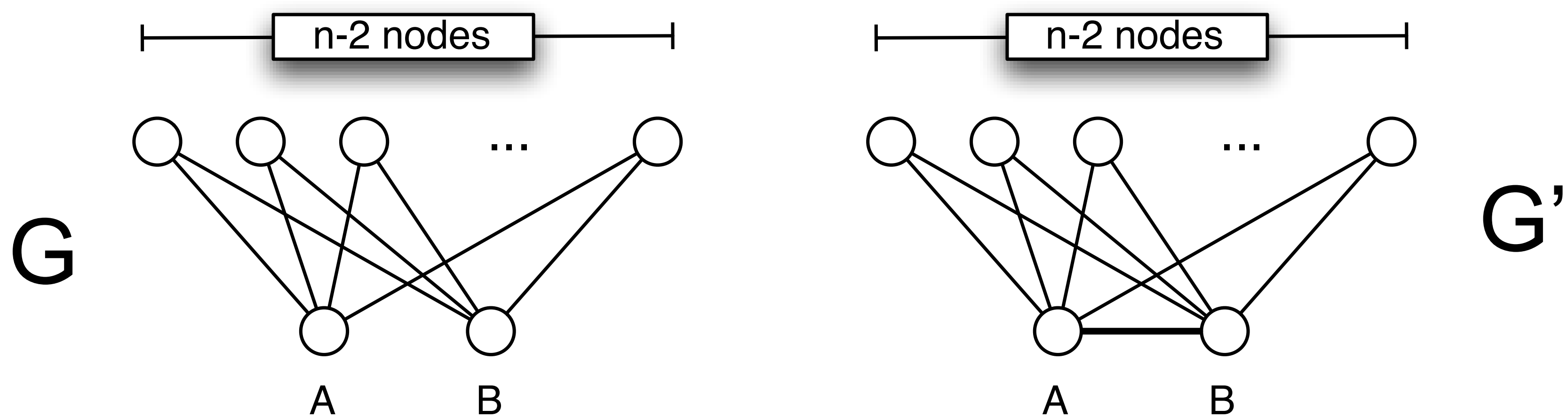


2-star **3-star** **triangle** **2-triangle**

- Importance
  - Used in statistical modeling: exponential random graph models
  - Descriptive statistics: clustering coefficient from 2-star, triangle

Subgraph counts have high (edge) sensitivity

- **Q_{TRIANGLE}**: return the number of triangles in the graph



Q_{TRIANGLE} (G) = 0

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

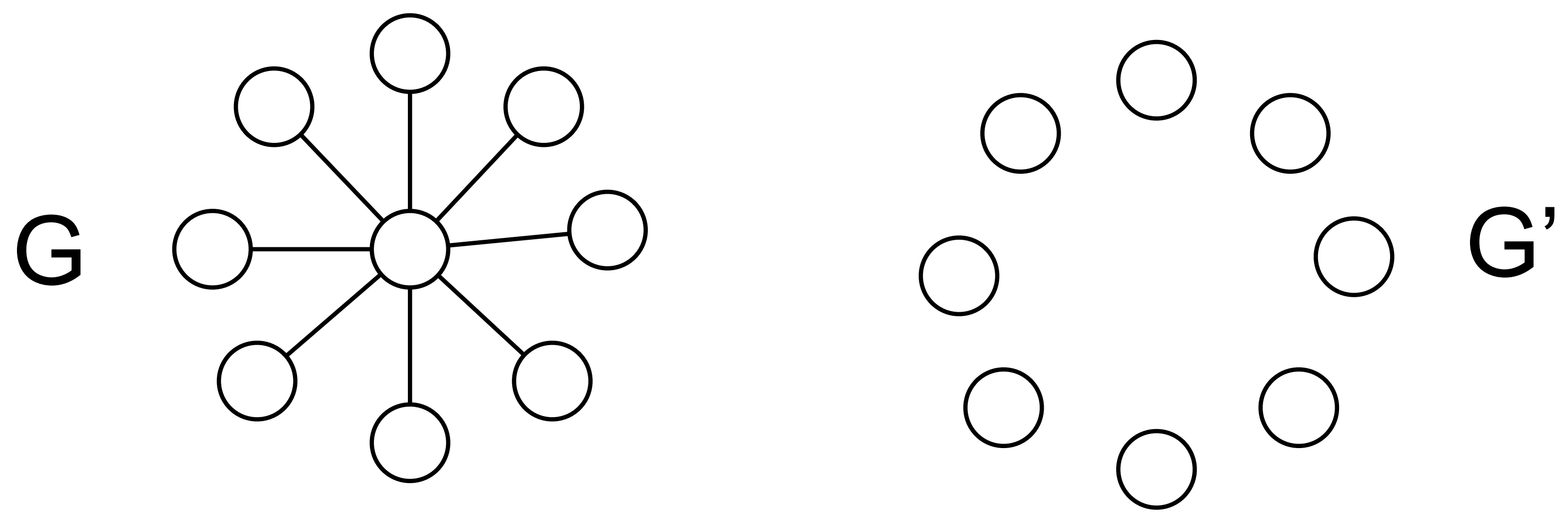
High Sensitivity:

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

- High sensitivity due "pathological" worst-case graph. If input is "far" from pathological, can we obtain accurate answers?

Degree queries have high (node) sensitivity

- **$Q_{DEGREE=d}$**: return the number of nodes of degree d in the graph



$Q_{DEGREE=1}$ (G) = 8 **$Q_{DEGREE=1}$ (G’) = 0**

High Sensitivity:

ΔQ_{DEGREE}=O(n)

- Every graph has a “pathological” neighbor. What accurate answers are possible?

Afternoon talk: <u>Sofya Raskhodnikova</u> “Survey of techniques for node-differential privacy”

Network analysis under **differential privacy**

- The **differential guarantee** for respondents in a data set:
  - Any information released about the sensitive data set must be virtually indistinguishable **whether or not a respondent's data is included in the dataset**.
- Sensitivity measures impact of changes to data
- Edge vs. node sensitivity