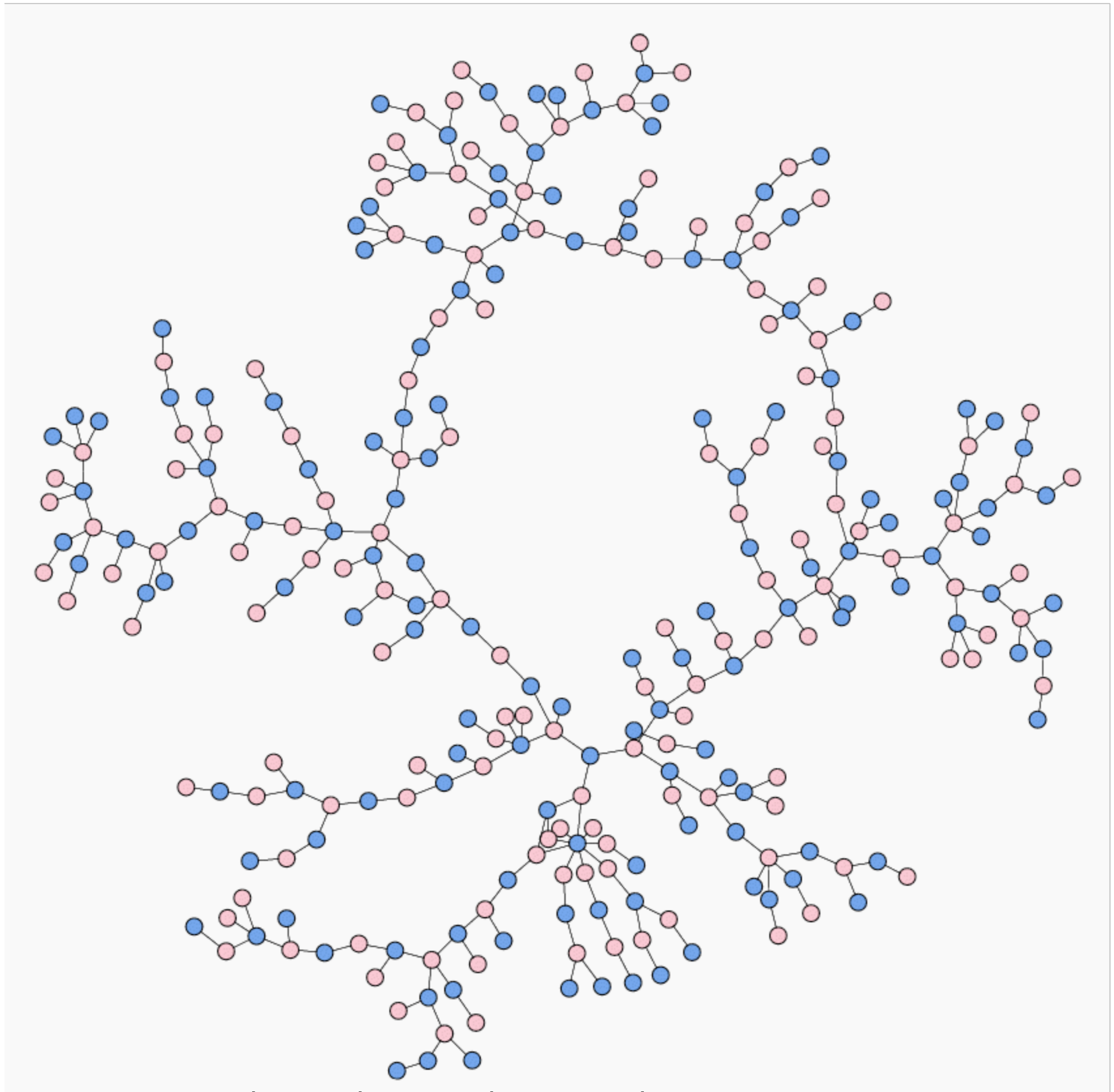
Privacy-Aware Data Management in Information Networks

Michael Hay, Cornell University
Kun Liu, Yahoo! Labs
Gerome Miklau, Univ. of Massachusetts Amherst
Jian Pei, Simon Fraser University
Evimaria Terzi, Boston University

SIGMOD 2011 Tutorial

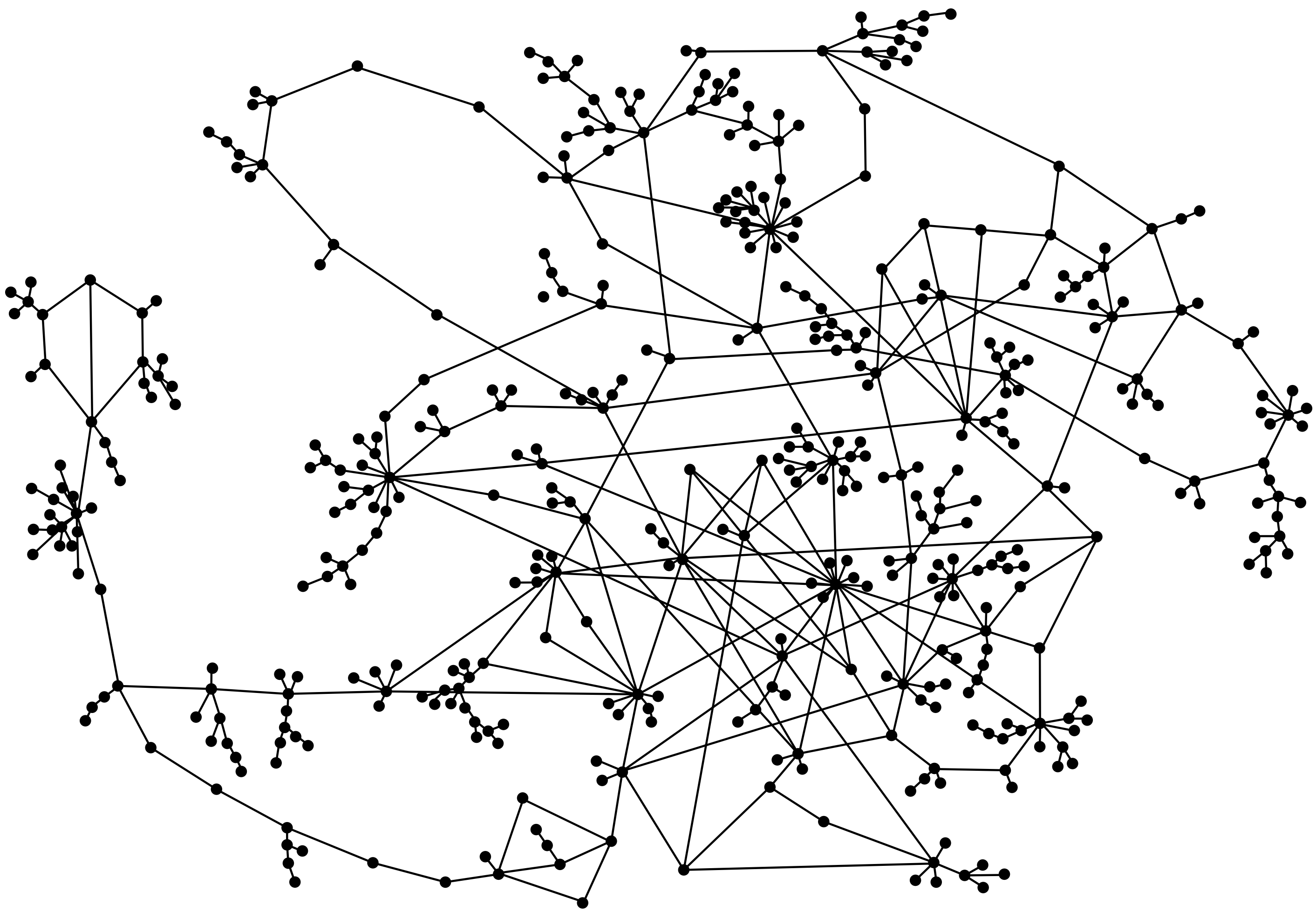
Romantic connections in a high school



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

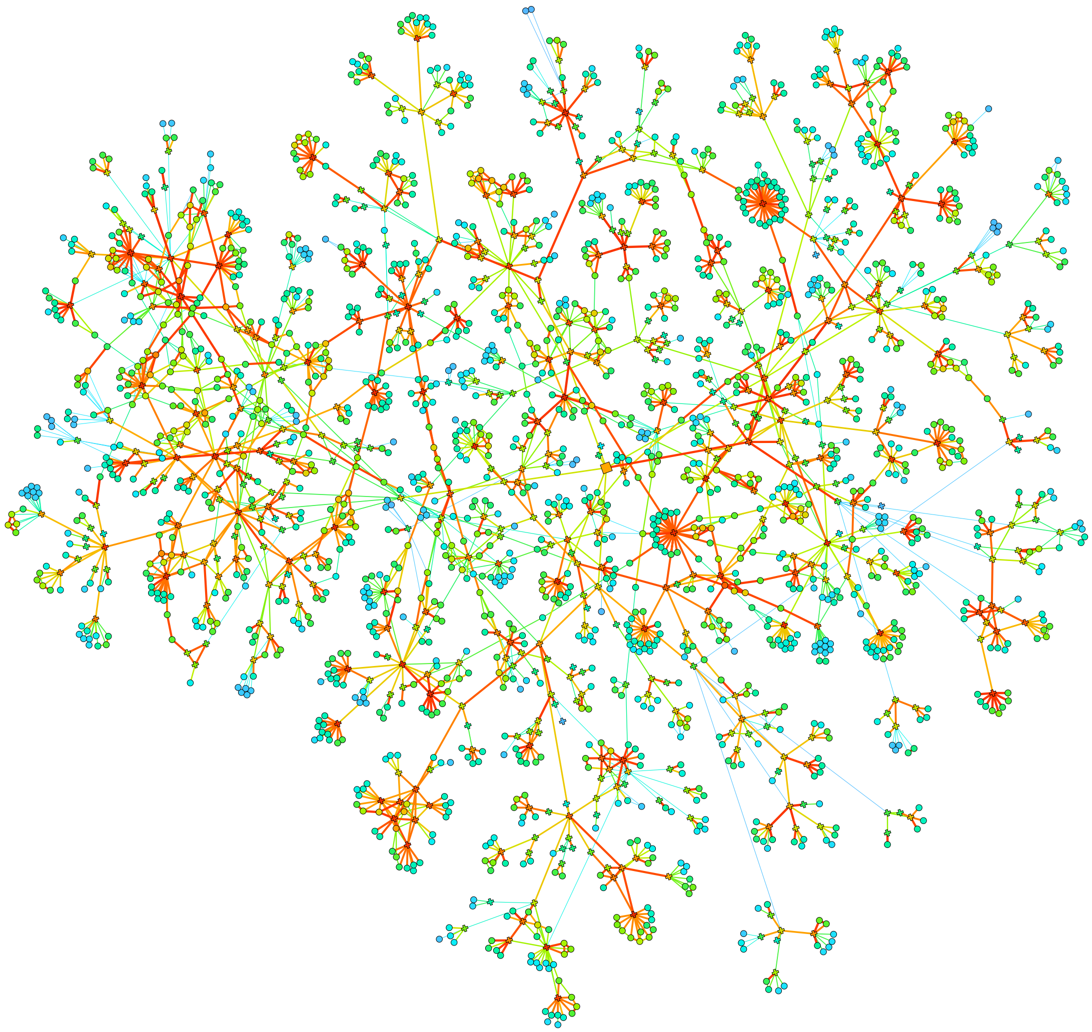
(Image drawn by Newman)

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*

Facebook

network data set

~600 million nodes
billions of relationships

complex platform for sharing

Privately managing enterprise network data	Personal Privacy in Online Social Networks
Data: Enterprise collects data or observes interactions of individuals.	**Data:** Individuals contribute their data thru participation in OSN.
Control: Enterprise controls dissemination of information.	**Control:** Individuals control their connections, interactions, visibility.
Goal: permit analysis of aggregate properties; protect facts about individuals.	**Goal:** reliable and transparent sharing of information.
Challenges: privacy for networked data, complex utility goals.	**Challenges:** system complexity, leaks thru inference, unskilled users.

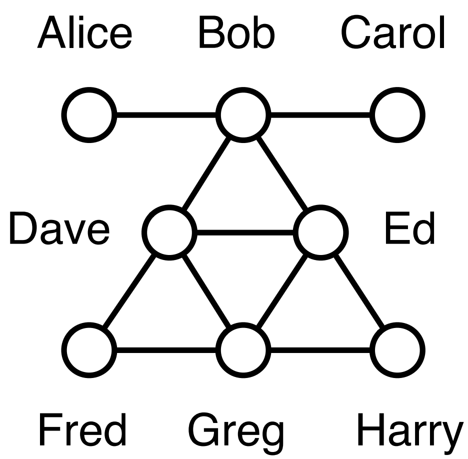
Outline of tutorial

- **Privately Managing Enterprise Network Data**
  - Goals, Threats, and Attacks
  - Releasing transformed networks (anonymity)
  - Releasing network statistics (differential privacy)
- **Personal Privacy in Online Social Networks**
  - Understanding privacy risk
  - Managing privacy controls

60 minutes

30 minutes

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 information in networks

- Disclosing attributes
- Disclosing edges
- Disclosing properties
    - node degree, clustering, etc.
    - properties of neighbors (e.g. mostly friends with republicans)

Goals in analyzing networks

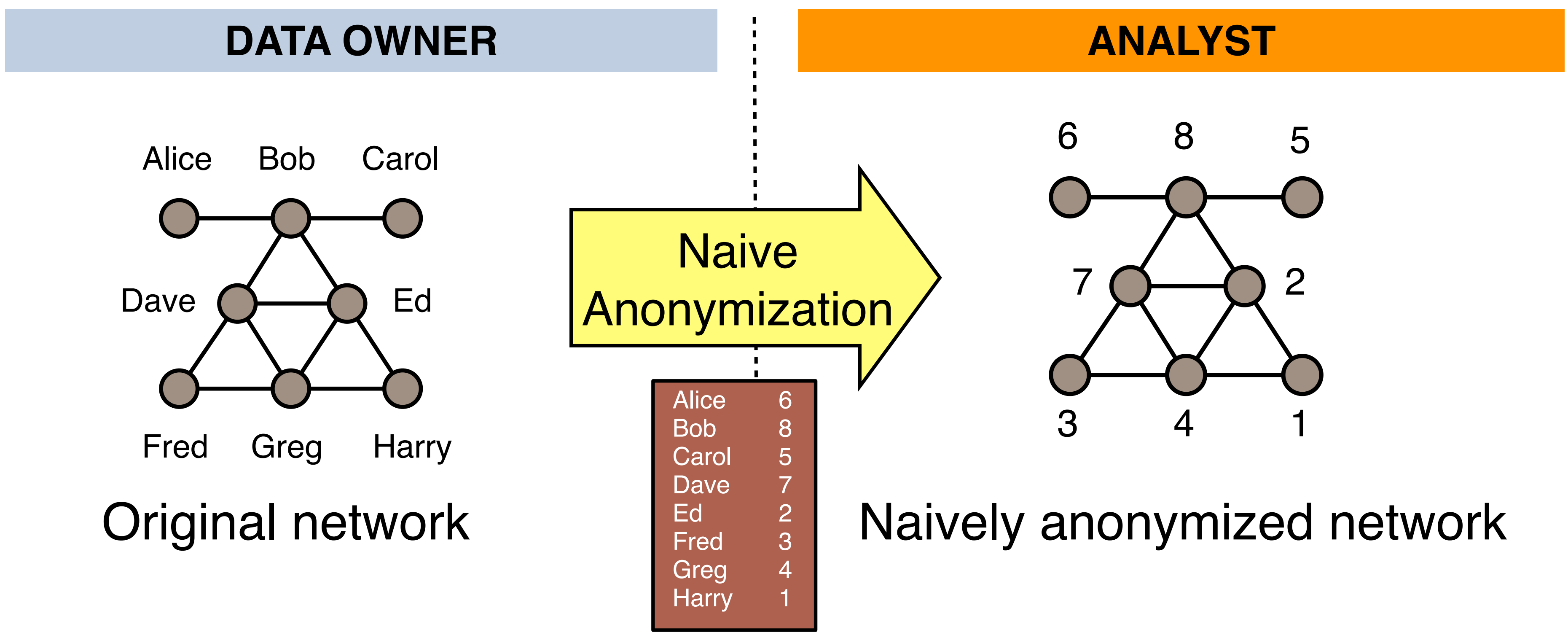
Can we permit analysts to study networks without revealing sensitive information about participants?

Example analyses

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

Naive anonymization

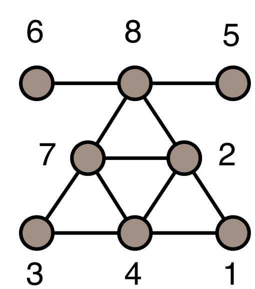
Naive anonymization is a transformation of the network in which identifiers are replaced with random numbers.



Good utility: output is isomorphic to the original network

Protection under naive anonymization

- Two primary threats:
  - **Node re-identification**: adversary is able to deduce that node x in the naively anonymized network corresponds to an identified individual Alice in the hidden network.
  - **Edge disclosure**: adversary is able to deduce that two identified individuals Alice and Bob are connected in the hidden network.
- With no external information: good protection
  - **Who is Alice?** **one of {1,2,3,4,5,6,7,8}**
  - **Alice and Bob connected?** **11/28 likelihood**



Adversaries with **external information**

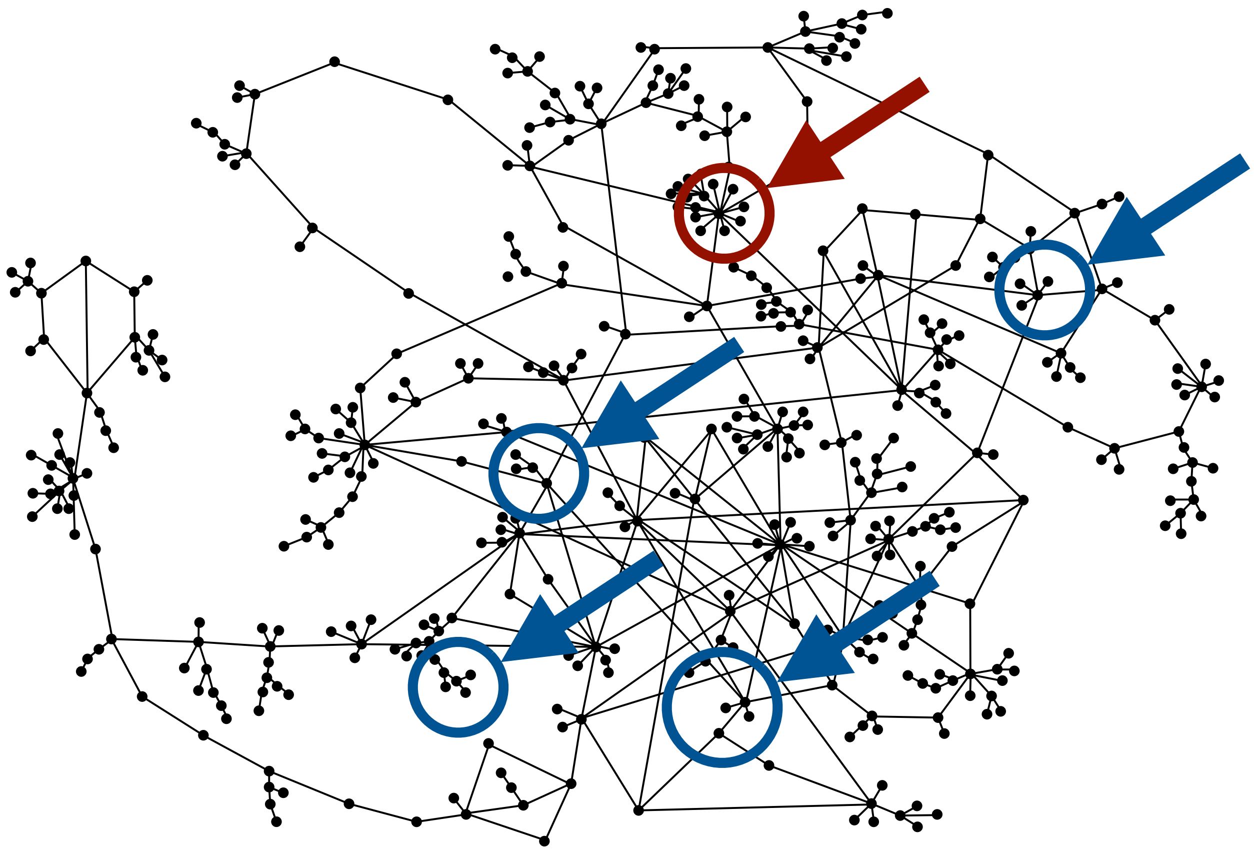
External information: facts about *identified* individuals and their relationships in the hidden network.

- Structural knowledge
  - often assumed limited to small radius around node
  - “Alice has degree 2” or “Bob has two connected neighbors”
- Information can be **precise** or **approximate**
- External information may be acquired from a **specific attack**, or we may assume a **category of knowledge** as a bound on adversary capabilities.

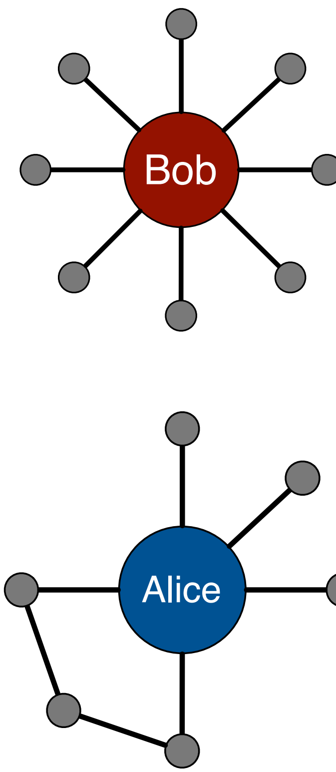
Matching attacks

Matching attack: the adversary matches external information to a naively anonymized network.

unique or **partial** **node re-identification**

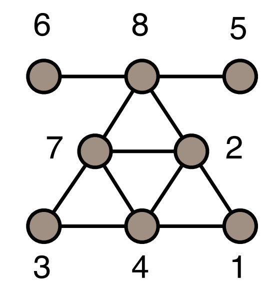


Naively Anonymized Network



External information

Attacks on naively anonymized networks

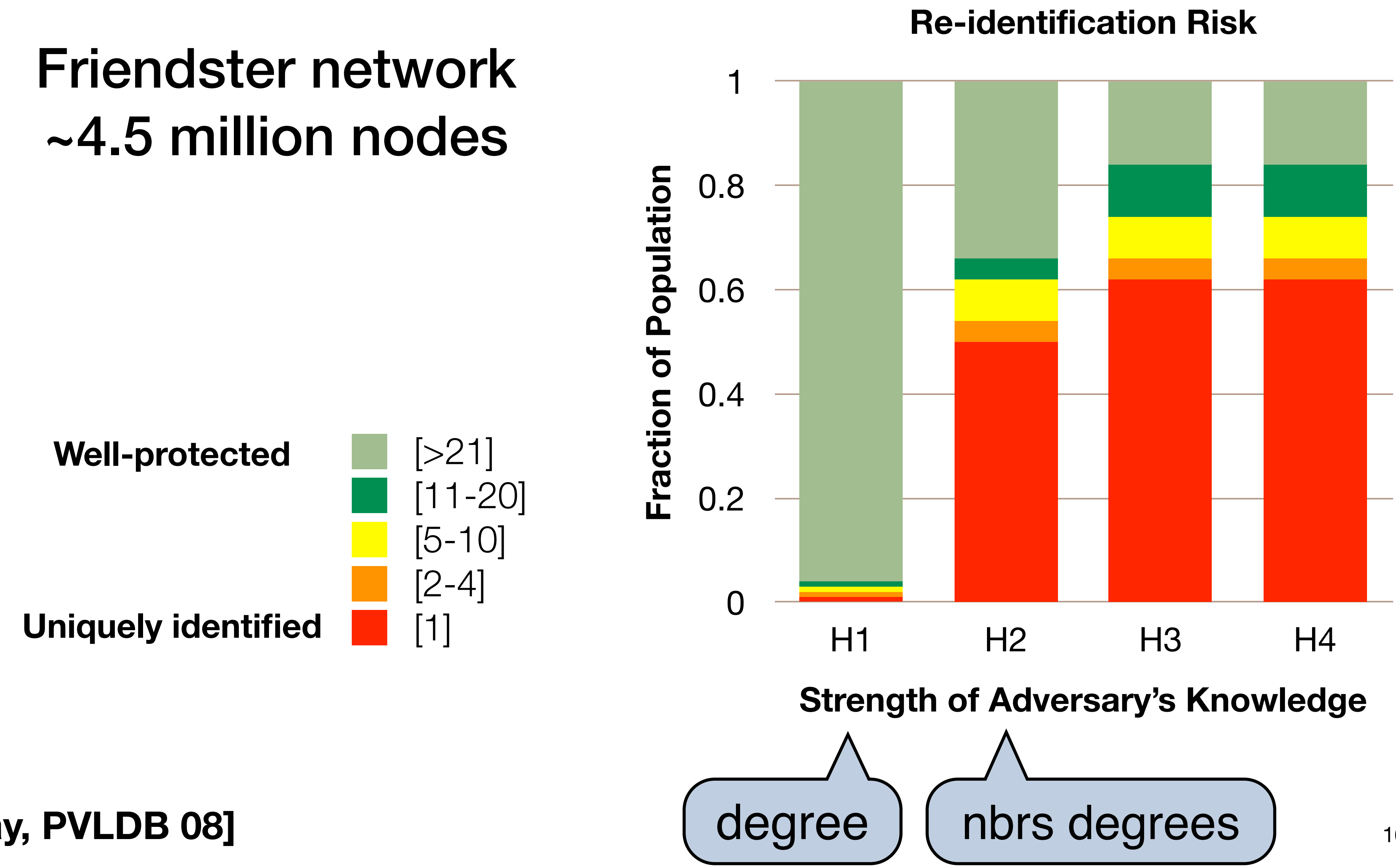


- Success of a matching attack depends on:
  - descriptiveness of external information
  - structural diversity in the network
- With external information: weaker protection
  - **Who is Alice?** **one of {1,2,3,4,5,6,7,8}**
  - **Who is Alice, if her degree is known to be 4 ?**

    one of {2,4,7,8}
  - **Alice and Bob connected?**

Local structure is highly identifying

Friendster network
~4.5 million nodes



[Hay, PVLDB 08]

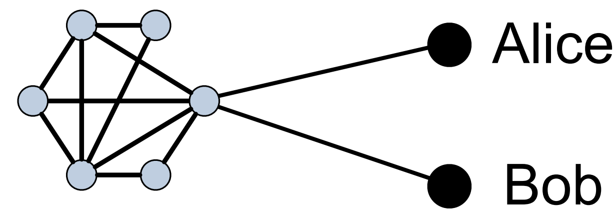
Active attack on an online network

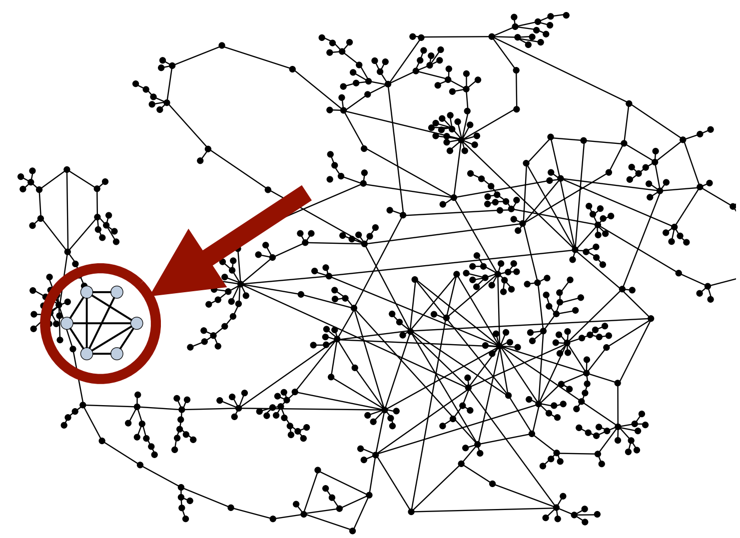
- Goal: **disclose edge** between two targeted individuals.
- Assumption: adversary can alter the network structure, by creating nodes and edges, **prior to** naive anonymization.
  - In blogging network: create new blogs and links to other blogs.
  - In email network: create new identities, send mail to identities.
  - (Harder to carry out this attack in a physical network)

[Backstrom, WWW 07]

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.





[Backstrom, WWW 07]

Results of active attack

- Given a network G with n nodes, it is possible to construct a pattern subgraph with $k = O(\log(n))$ nodes that will be unique in G with high probability.
    - injected subgraph is chosen uniformly at random.
    - the number of subgraphs of size k that appear in G is small relative to the number of all possible subgraphs of size k.
- The pattern subgraph can be efficiently found in the released network, and can be linked to as many as $O(\log^2(n))$ target nodes.
- In 4.4 million node Livejournal friendship network, attack succeeds w.h.p. for 7 pattern nodes.

[Backstrom, WWW 07]

Auxiliary network attack

- Goal: re-identify individuals in a naively anonymized target network
- Assumptions:
  - An un-anonymized auxiliary network exists, with overlapping membership.
  - There is a set of seed nodes present in both networks, for which the mapping between target and auxiliary is known.
- Starting from seeds, mapping is extended greedily.
- Using Twitter (target) and Flickr (auxiliary), true overlap of ~30000 individuals, 150 seeds, 31% re-identified correctly, 12% incorrectly.

[Narayanan, OAKL 09]

Summary

- Naive anonymization may be good for utility...
- ... but it is **not sufficient** for protecting sensitive information in networks.
  - an individual's connections in the network can be highly identifying.
  - external information may be available to adversary from outside sources or from specific attacks.
- Conclusion: stronger protection mechanisms are required.

Questions & challenges

- What is the correct model for adversary external information?
- How do attributes and structural properties combine to increase identifiability and worsen attacks?
- Are there additional attacks on naive anonymization (or other forms of anonymization)?

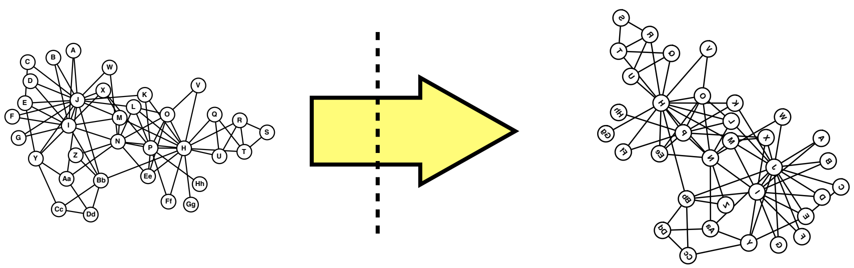
Next: How can we strengthen the protection offered by a released network while preserving utility ?

Outline of tutorial

- **Privately Managing Enterprise Network Data**
  - Goals, Threats, and Attacks
  - Releasing transformed networks (anonymity)
  - Releasing network statistics (differential privacy)
- **Personal Privacy in Online Social Networks**
  - Understanding privacy risk
  - Managing privacy controls

Releasing data vs. statistics

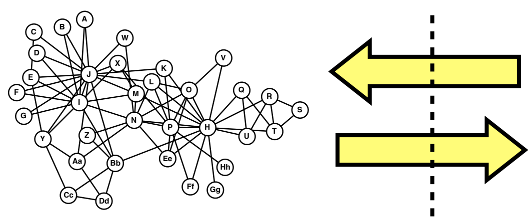
- **Releasing transformed networks**



To prevent adversary attack, release *transformed* network

- transformations obscure identifying node features
- while hopefully preserve global topology

- **Releasing “safe” network statistics**

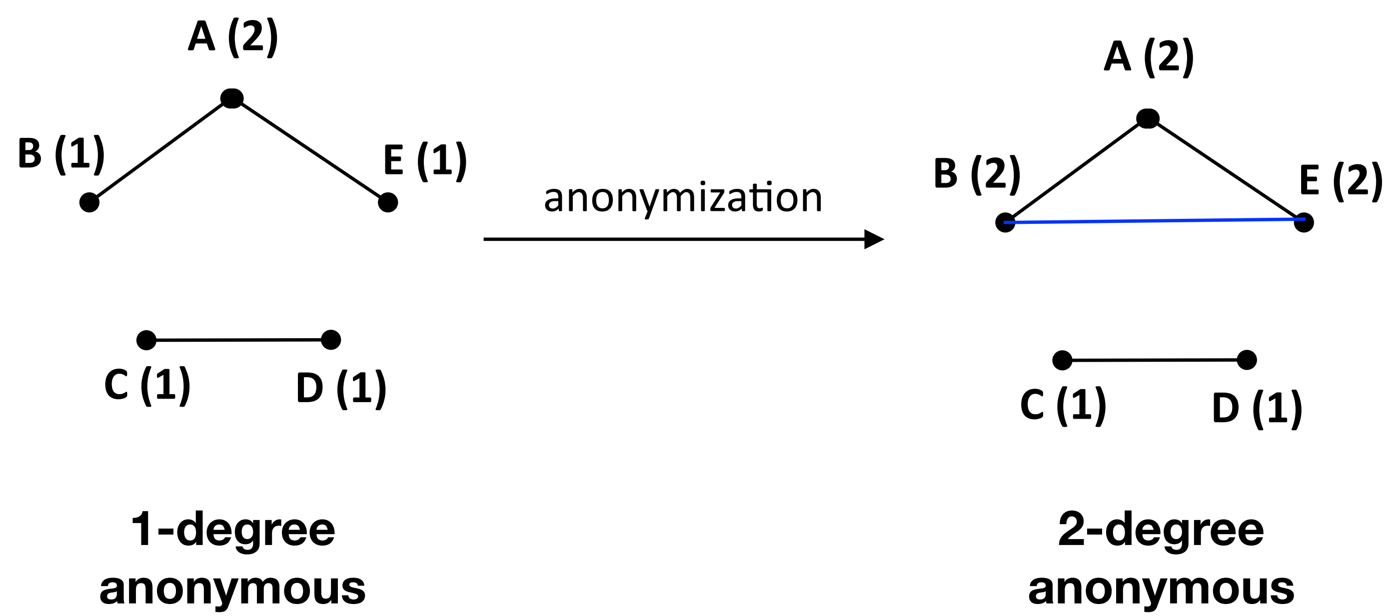


Q

safe answer

Transform for degree anonymity

- A graph $G(V, E)$ is k-degree anonymous if every node in V has the same degree as k-1 other nodes in V.



[Liu, SIGMOD 08]

[Liu, SIGMOD 08]

Algorithm for degree anonymization

- Problem: Given a graph *G*(*V*, *E*) and integer *k*, find minimal set of edges *E'* such that graph *G*(*V*, *E* ∪ *E*') is *k*-degree anonymous.
- Approach: Use dynamic programming to finds minimum change to degree sequence.
- Challenge: may not be possible to *realize* degree sequence through edge additions.
- Example: *V* = {*a, b, c*}, *E* = { (*b,c*) }. Degree sequence is [0,1,1]. Min. change yields [1,1,1] but not realizable (without self-loops).
- Algorithm: draws on ideas from graph theory to construct a graph with minimum, or near minimum, edge insertions.

A common problem formulation

- Degree anonymization is an instance of a more general paradigm. Many approaches proposed follow this paradigm.

Given input graph G,

- Consider set of graphs $\mathcal{G}$, each G^* in $\mathcal{G}$ reachable from G by certain graph **transformations**
- Find G^* in $\mathcal{G}$ such that G^* satisfies **privacy**($G^*, \ldots$), and
- Minimizes **distortion**(G, G^*)

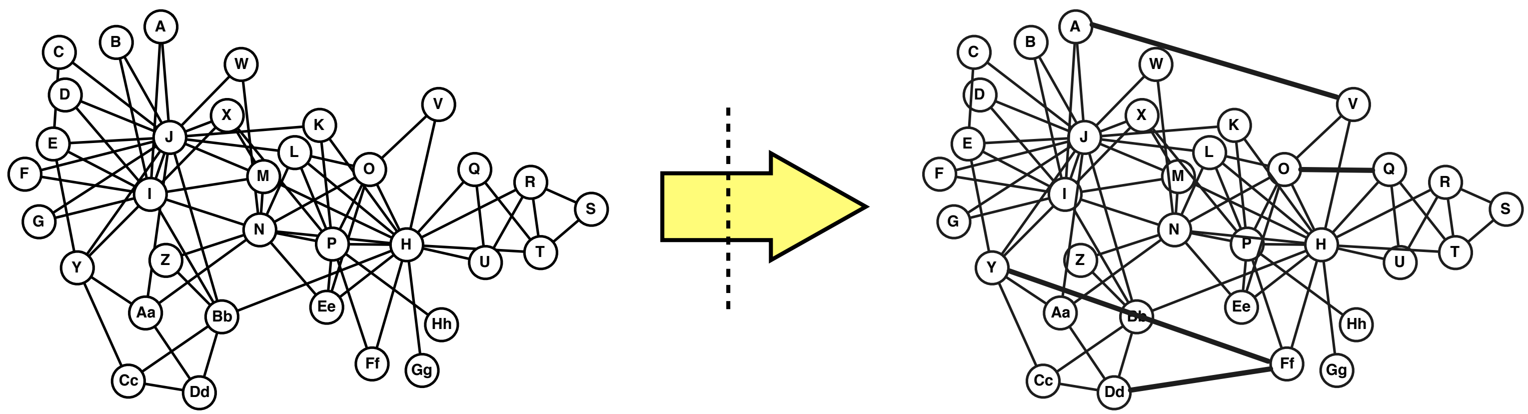
Privacy as resistance to attack

- Adversary capability: knowledge of...
    - attributes
    - degree
    - subgraph neighborhood
    - structural knowledge beyond immediate neighborhood
- Attack outcome
    - Node re-identification
    - Edge disclosure

Kinds of transformations

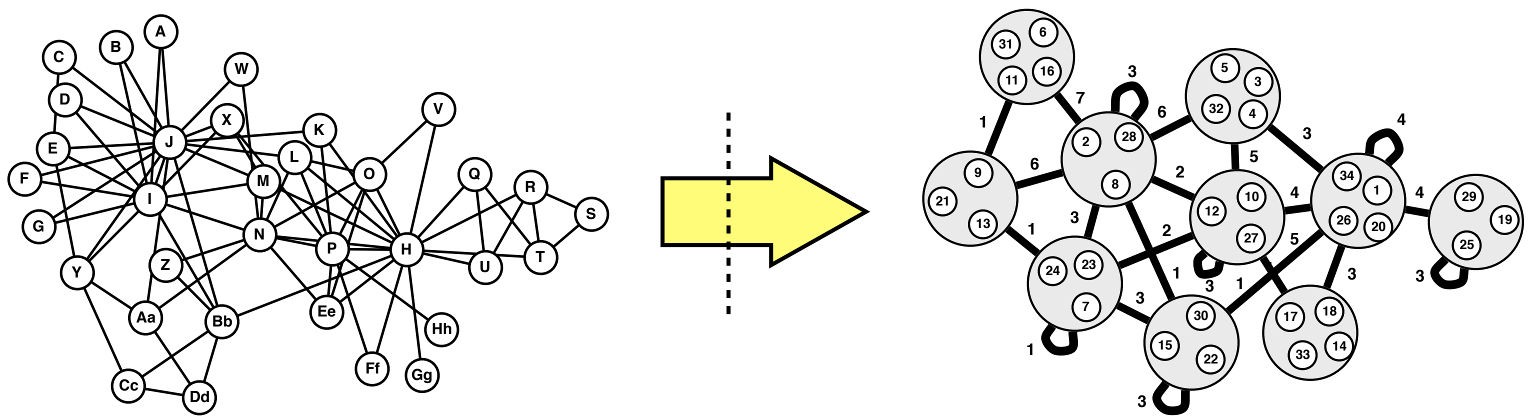
- Transformations considered in literature can be classified into three categories
  - Directed alteration
  - Generalization
  - Random alteration

Directed alteration



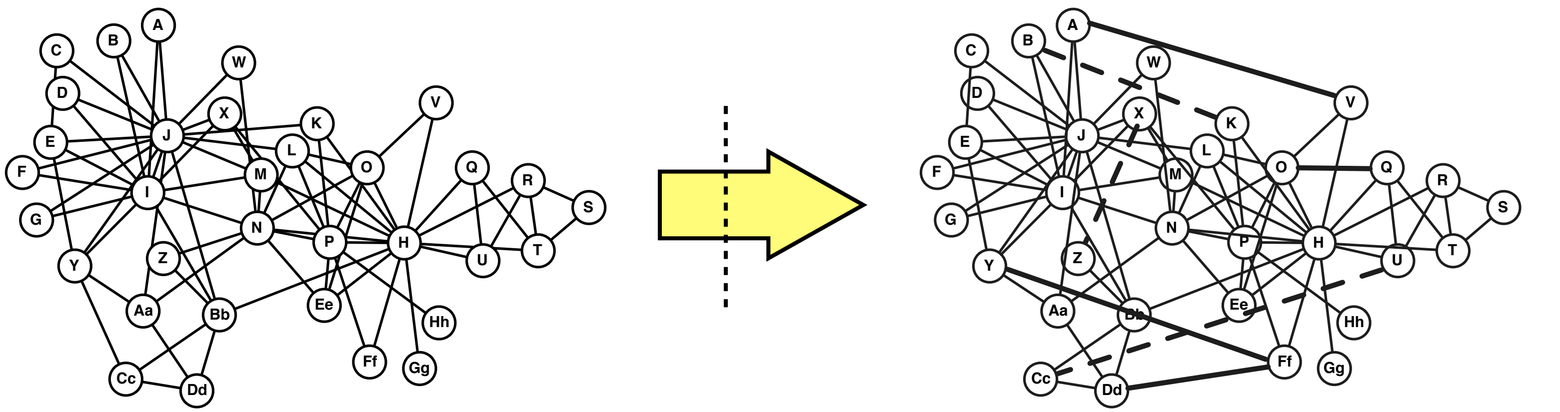
- Transform network by adding (or removing) edges
  - **[Liu, SIGMOD 08]** insert edges to achieve degree anonymity
  - **[Zhou, ICDE 08]** neighborhood anonymity, labels on nodes
  - **[Zou, PVLDB 09]** complete anonymity (k isomorphic subgraphs)
  - **[Cheng, SIGMOD 10]** complete anonymity and bounds on edge disclosure

Generalization



- Transform network by cluster nodes into groups
  - **[Cormode, PVLDB 08]** attribute-based attacks (graph structure unmodified) on bipartite graphs, prevents edge disclosure
  - **[Cormode, PVLDB 09]** similar to above but for arbitrary interaction graphs (attributes on nodes and edges)
  - **[Hay, PVLDB 08, VLDBJ 10]** summarize graph topology in terms of node groups; anonymity against arbitrary structural knowledge

Random alteration



- Transform network by stochastically adding, removing, or rewiring edges
  - **[Ying, SDM 08]** random rewiring subject to utility constraint (spectral properties of graph must be preserved).
  - **[Liu, SDM 09]** randomization to hide sensitive edge weights
  - **[Wu, SDM 10]** exploits spectral properties of graph data to filter out some of the introduced noise.

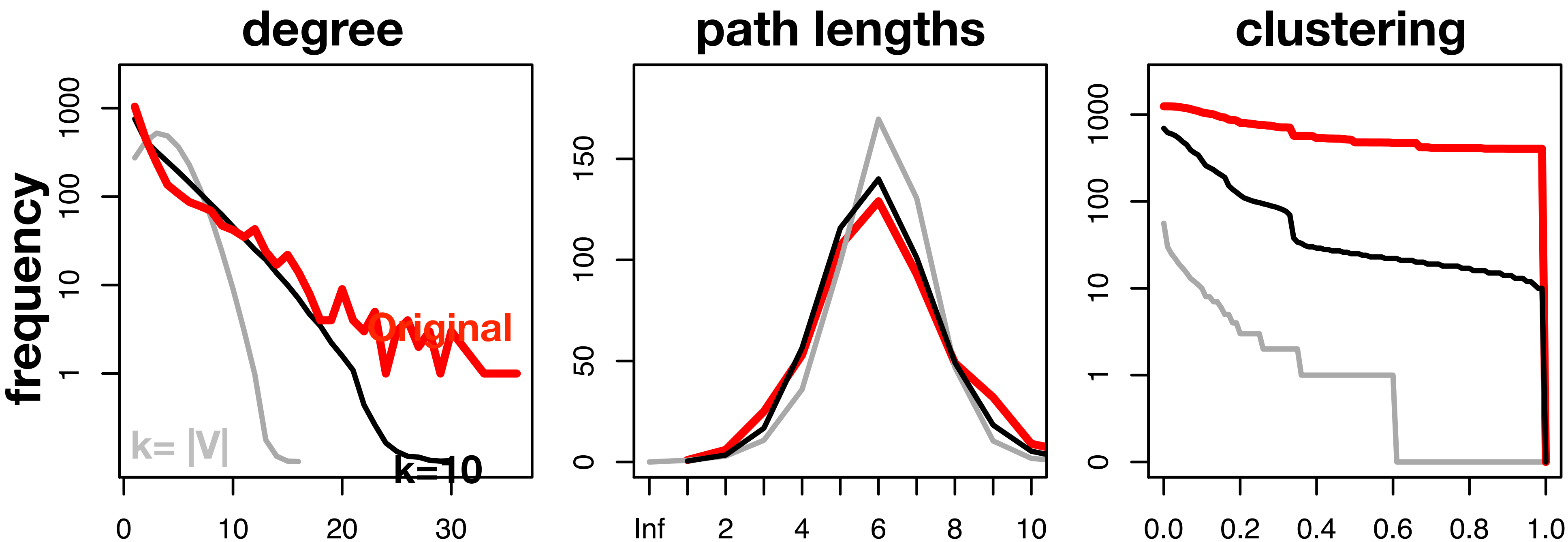
Other work in network transformation

- Other works
  - **[Zheleva, PinKDD 07]** predicting sensitive hidden edges from released graph data (nodes and non-sensitive edges).
  - **[Ying, SNA-KDD 09]** comparison of randomized alteration and directed alteration.
  - **[Bhagat, WWW 10]** releasing multiple views of a dynamic social network.
- Surveys:
  - **[Liu, Next Generation Data Mining 08]**
  - **[Zhou, SIGKDD 08]**
  - **[Hay, Privacy-Aware Knowledge Discovery 10]**
  - **[Wu, Managing and Mining Graph Data 10]**

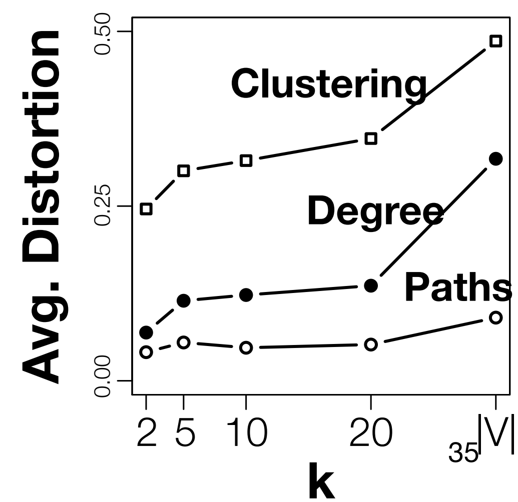
Evaluating impact on utility

- After transformations, graph is released to public. Analyst measures transformed graph in place of original. What is impact on utility?
- Graph remains useful if it is "similar" to original. How measure similarity?
- Related questions arise in statistical modeling of networks and assessing model fitness **[Goldenberg, Foundations 10] [Hunter, JASA 08]**
- Common approach to evaluating utility: empirically compare transformed graph to original graph in terms of various network properties

Impact on network properties



Algorithm from Hay PVLDB 08; experiments on version of HepTh network (2.5K nodes, 4.7K edges)



[Hay, PVLDB 08]

Limitations

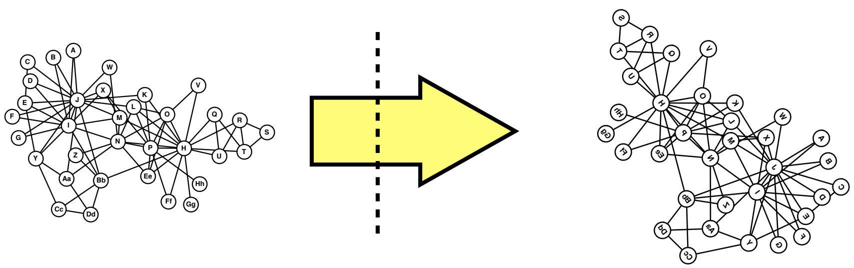
- Utility
    - Transformation may distort some properties: some analysts will find transformed graph useless
    - Lack of formal bounds on error: analyst *uncertain* about utility
- Privacy
    - Defined as resistance to a *specific* class of attacks; vulnerable to unanticipated attacks?
    - Inspired by k-anonymity; doomed to repeat that history? (See survey **[Chen, Foundations and Trends in Database 09].**)

Outline of tutorial

- **Privately Managing Enterprise Network Data**
  - Goals, Threats, and Attacks
  - Releasing transformed networks (anonymity)
  - Releasing network statistics (differential privacy)
    - Differential privacy
    - Degree sequence
    - Subgraph counts
- **Personal Privacy in Online Social Networks**

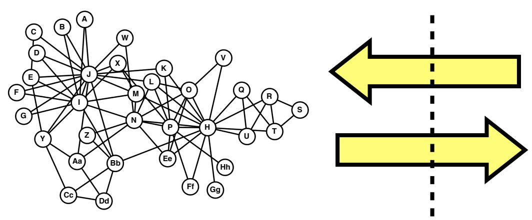
Releasing data vs. statistics

- **Releasing transformed networks**



Ease of use	good
Protection	anonymity
Accuracy	no formal guarantees

- **Releasing "safe" network statistics**



Q

safe answer

↓

output perturbation

Q(G) + noise

Ease of use	bad for practical analyses
Protection	formal privacy guarantee
Accuracy	provable bounds

When are aggregate statistics safe to release?

- “Safe” statistics should report on properties of a group, without revealing properties of individuals.
    - We often want to release a combination of statistics. Still safe?
    - What if adversary uses external information along with statistics? Still safe?
- Dwork, McSherry, Nissim, Smith **[Dwork, TCC 06]** proposed **differential privacy** as a rigorous standard for safe release.
- Many existing results for tabular data; relatively few results for network data.

The differential guarantee

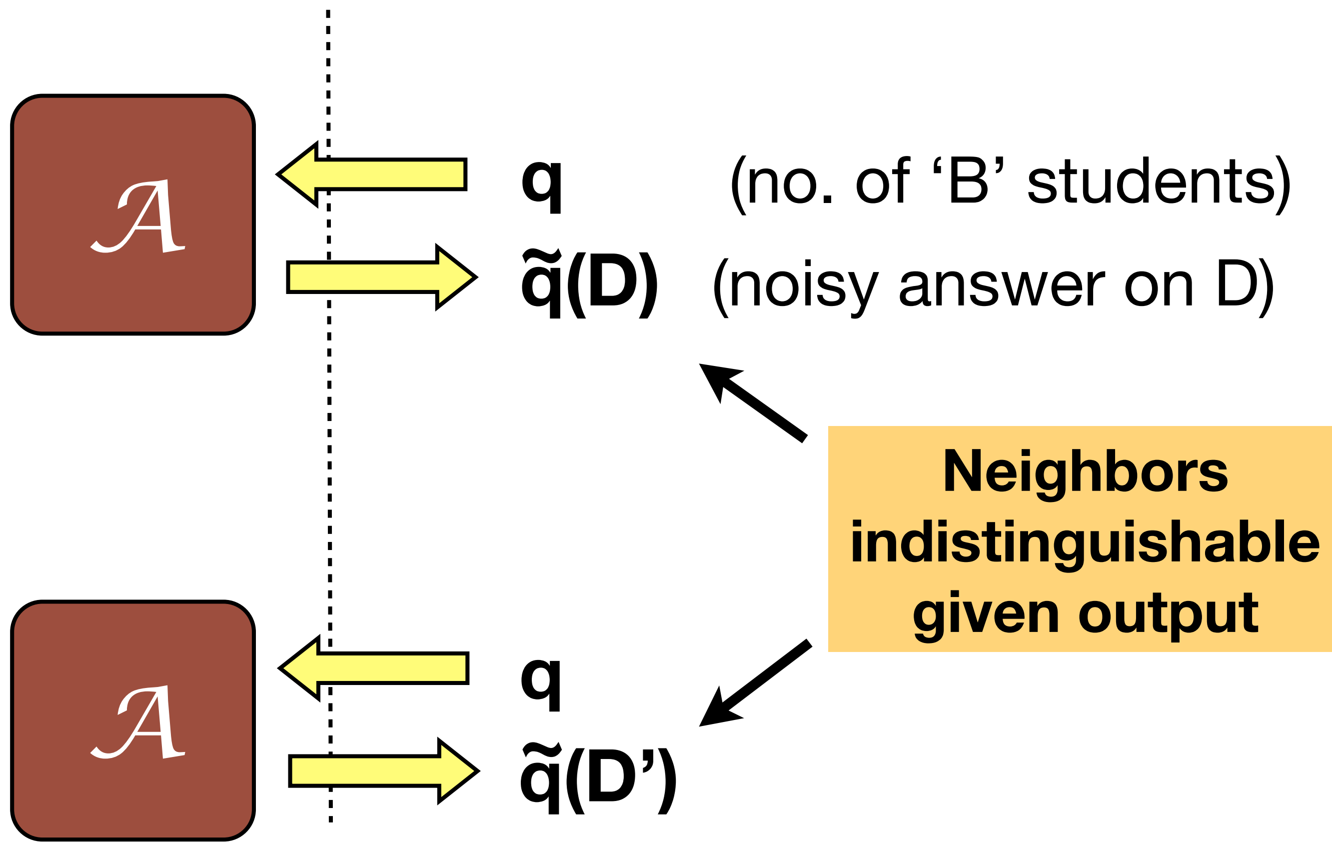
DATA OWNER | **ANALYST**

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

Differential privacy

A randomized algorithm A provides **Ɛ-differential privacy** if:
for all neighboring databases D and D', and
for any set of outputs *S*:

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

epsilon is a privacy parameter

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

$\Downarrow$ epsilon = $\Uparrow$ stronger privacy

[Dwork, TCC 06]

Calibrating noise

- How much noise is necessary to ensure differential privacy?
- Noise large enough to hide “contribution” of individual record.
- Contribution measured in terms of query **sensitivity.**

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

Query q	Sensitivity Δq
q1: Count tuples	1
q2: Count('B' students)	1
q3: Count(students with property X)	1
q4: Median(age of students)	~ max age

[Dwork, TCC 06]

[Dwork, TCC 06]

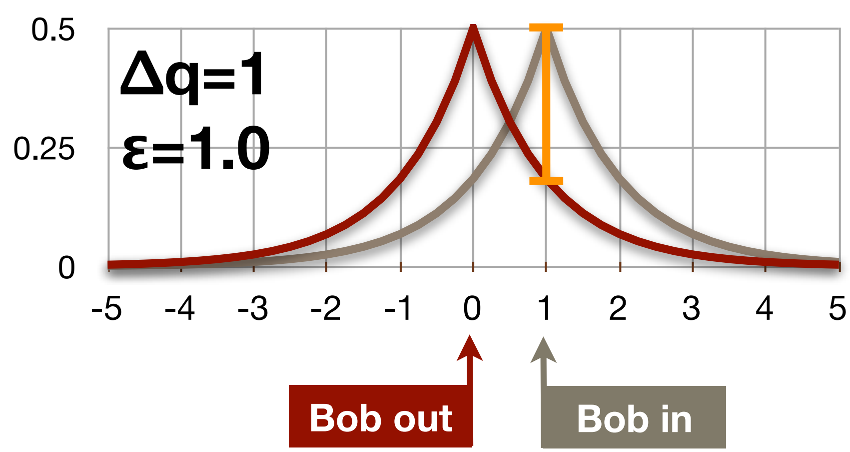
The Laplace mechanism

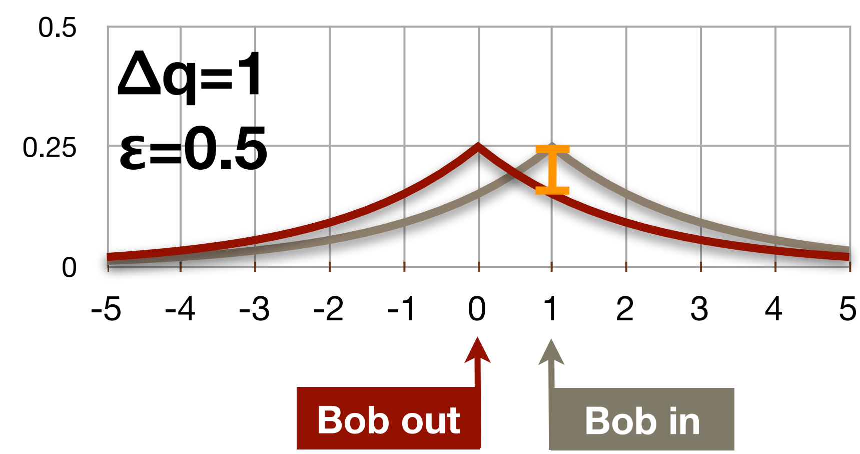
The following algorithm for answering **q** is ε-differentially private:

sensitivity of **q**

privacy parameter

$\mathcal{A}$ **Laplace Mechanism** → **q(D) + Laplace(Δq / ε)**





Differentially private algorithms

- Any query can be answered (but perhaps with lots of noise)
- Noise determined by privacy parameter epsilon and the sensitivity (both public)
- Multiple queries can be answered (details omitted)
- Privacy guarantee does not depend on assumptions about the adversary (caveats omitted, see **[Kifer, SIGMOD 11]**)

Survey paper on differential privacy: **[Dwork, CACM 10]**

Adapting differential privacy for networks

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

- For networks, what is the right notion of "differential object?"
  - Hide individual's "evidence of participation" **[Kifer, SIGMOD 11]**
  - An edge? A set of *k* edges? A node (and incident edges)?
  - More discussion in **[Hay, ICDM 09] [Kifer, SIGMOD 11]**
- Choice impacts utility
- Existing work considers only edge, and k-edge, differential privacy.

What can we learn accurately?

- What can we learn accurately about a network under edge or k-edge differential privacy?
- Basic approach:
  - Express desired task as one or more queries.
  - Check query sensitivity
    - **if High**: not promising, but sometimes representation matters.
    - **if Low:** maybe promising, but may still require work.

Outline of tutorial

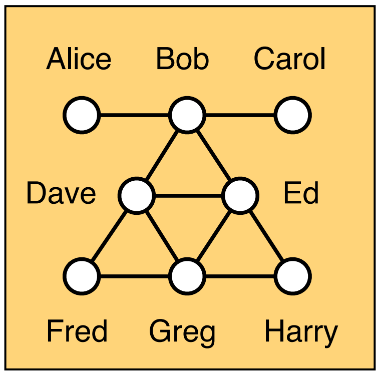
- **Privately Managing Enterprise Network Data**
  - Goals, Threats, and Attacks
  - Releasing transformed networks (anonymity)
  - Releasing network statistics (differential privacy)
    - Differential privacy
    - Degree sequence
    - Subgraph counts
- **Personal Privacy in Online Social Networks**

[Hay, ICDM 09] [Hay, PVLDB 10]

The degree sequence can be estimated accurately

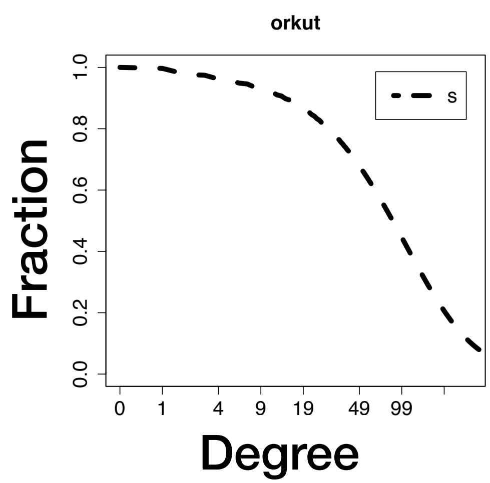
- Degree sequence: the list of degrees of each node in a graph.

- A widely studied property of networks.



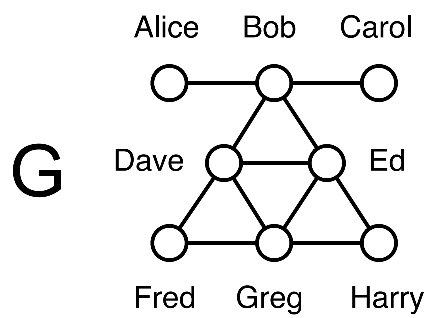
[1,1,2,2,4,4,4,4]

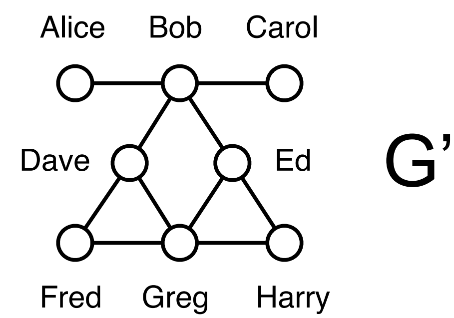
Orkut crawl



Inverse cumulative distribution

Two basic queries for degrees





Degree of each node	
deg_A	degree of node A
D	$[deg_A, deg_B, \ldots \quad]$

Frequency of each degree	
cnt_i	count of nodes with degree i
F	$[cnt_0, cnt_1, \ldots \quad cnt_{n-1}]$

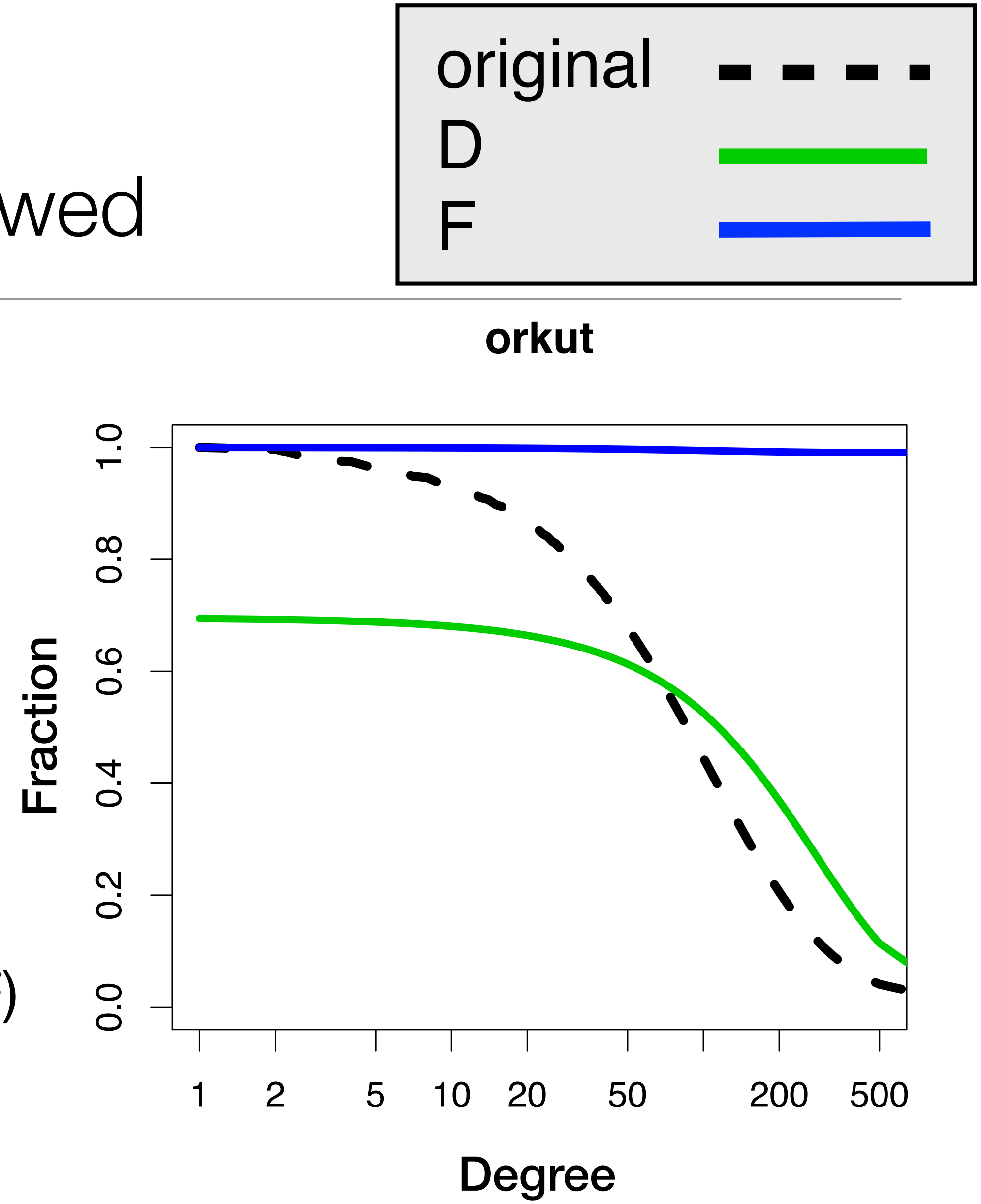
$D(G) = [1,4,1,4,4,2,4,2]$
$D(G') = [1,4,1,\underline{3},\underline{3},2,4,2]$

$\boxed{\Delta D=2}$

$F(G) = [0,2,2,0,4,0,0,0]$
$F(G') = [0,2,2,\underline{2},\underline{2},0,0,0]$

$\boxed{\Delta F=4}$

These queries are both flawed



- D requires independent samples from Laplace(2/ε) in each component.
- F requires independent samples from Laplace(4/ε) in each component.
- Thus Mean Squared Error is $\Theta(n/\varepsilon^2)$

New technique allows improved error of $O(d \log^3(n)/\varepsilon^2)$
(where d is # of unique degrees)

[Hay, ICDM 09] [Hay, PVLDB 10]

An alternative query for degrees

Degree of each node	
deg_A	degree of node A
D	$[deg_A, deg_B, \ldots \quad]$

Degree of each node, ranked	
rnk_i	return the rank i^{th} degree
S	$[rnk_1, rnk_2, \ldots \quad rnk_n]$

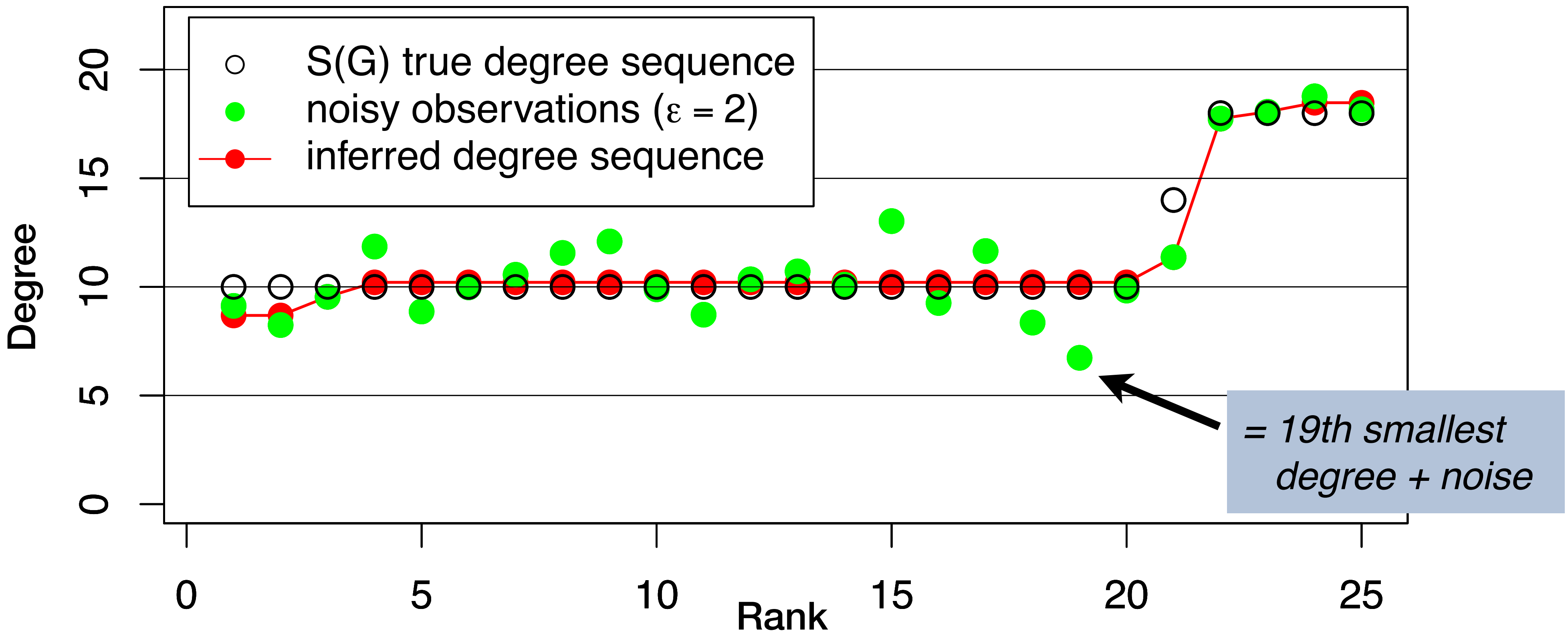
D(G) = [1,4,1,4,4,2,4,2]
D(G’) = [1,4,1,$\underline{3}$,$\underline{3}$,2,4,2]

S(G) = [1,1,2,2,4,4,4,4]
S(G’) = [1,1,2,2,$\underline{3}$,$\underline{3}$,4,4]

$\Delta D=2$

$\Delta S=2$

Using the sort constraint

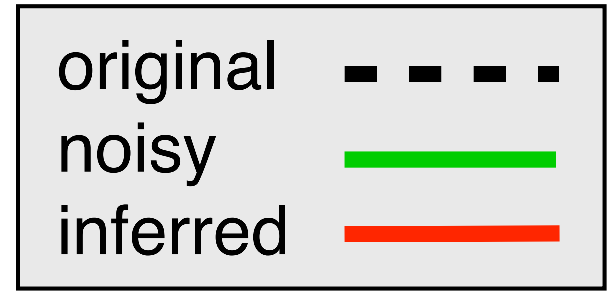


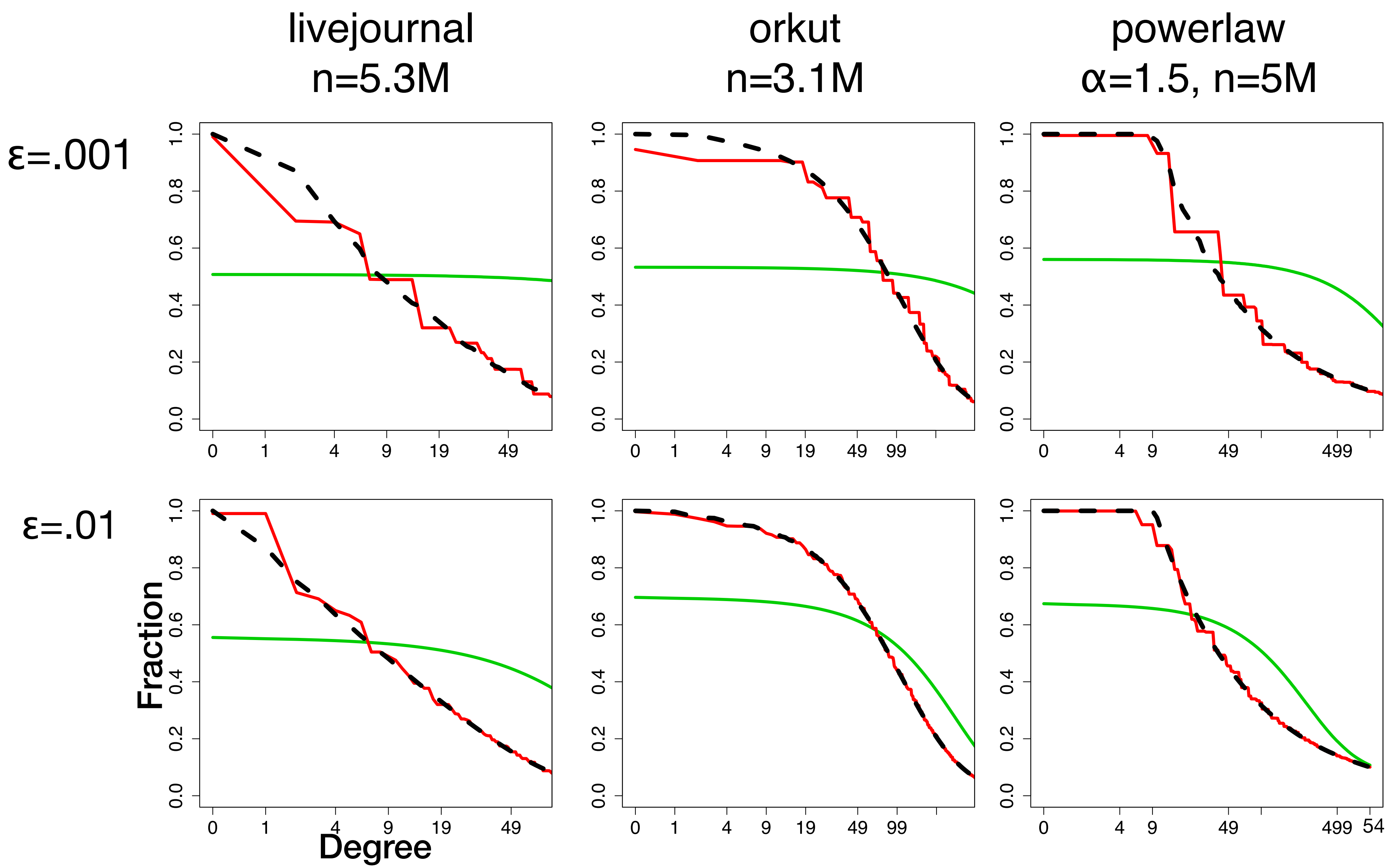
- The output of the sorted degree query is not (in general) sorted.

S(G) = [10, 10, ...10, 10, 14, 18,18,18,18]

- We derive a new sequence by computing the **closest** non-decreasing sequence: i.e. minimizing L2 distance.

Experimental results, continued



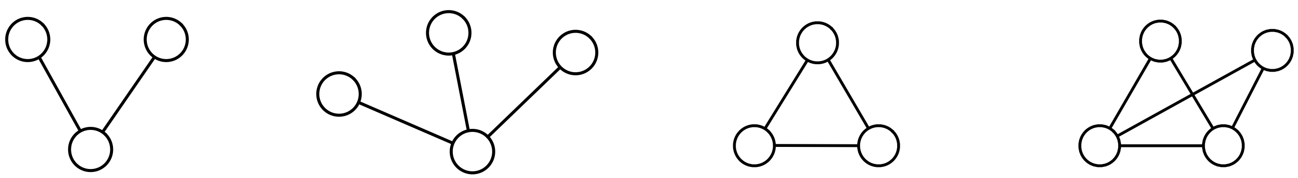


Outline of tutorial

- **Privately Managing Enterprise Network Data**
  - Goals, Threats, and Attacks
  - Releasing transformed networks (anonymity)
  - Releasing network statistics (differential privacy)
    - Differential privacy
    - Degree sequence
    - Subgraph counts
- **Personal Privacy in Online Social Networks**

Subgraph counting queries

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



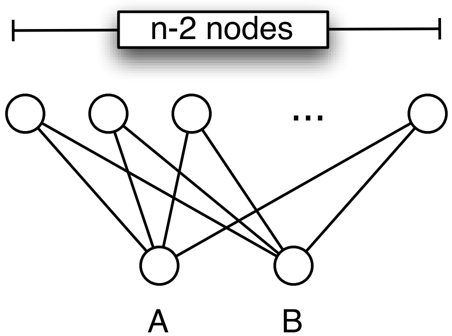
2-star **3-star** **triangle** **2-triangle**

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

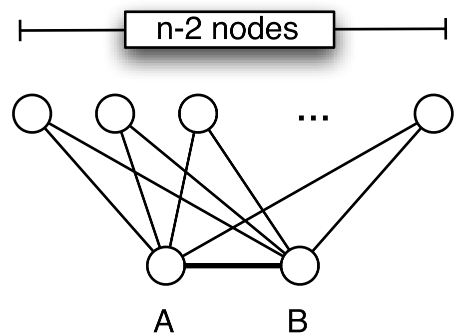
Subgraph counts have high sensitivity

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

G



$Q_{TRIANGLE}$ (G) = 0



G'

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

High Sensitivity:

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

- High sensitivity due "pathological" worst-case graph. If input is not pathological, can we obtain accurate answers?

Local sensitivity

- Tempting, but flawed, idea is to add noise proportional to **local** sensitivity.

- **Local** sensitivity of q on G: maximum difference in query answer between G and a neighbor G'.

$$LS(G) = \max_{G' \in N(G)} | q(G) - q(G') |$$

- Example shows problem of using local sensitivity (from **[Smith, IPAM 10]**): database D is set of number, query q is the median

LS(D)=0

$$D = \underbrace{0 \ldots 0}_{(n-3)/2}\ 000\ \underbrace{c \ldots c}_{(n-3)/2}$$

LS(D')=c

$$D' = \underbrace{0 \ldots 0}_{(n-3)/2}\ 00\underline{c}\ \underbrace{c \ldots c}_{(n-3)/2}$$

Instance-based noise

- Two general approaches to adding instance-based noise
  - **Smooth sensitivity** Compute a *smooth* upper bound on local sensitivity **[Nissim, STOC 07].**
  - **Noisy sensitivity** Use differentially private mechanism to get noisy upper “bound” on local sensitivity **[Behoora, PVLDB 11] [Dwork, STOC 09].**
- Instance-based noise can require modest relaxation of differential privacy to account for (very low probability) “bad” events.

Differentially private subgraph counts

- For *k*-stars and triangles, *smooth sensitivity* is efficiently computable
- For *k*-triangles with $k \geq 2$
  - Computing *smooth sensitivity* NP-Hard.
  - However, it can be estimated using *noisy sensitivity* approach
- Empirical and theoretical analysis:
  - Generally, instance-based noise not much larger than local sensitivity
  - However, for k-triangles on real data, local sensitivity sometimes large (relative to actual number of k-triangles).

[Behoora, PVLDB 11]

Alternative representations

- Number of k-stars in a graph can be computed from the degree sequence

$$\text{k-stars}(G) = \sum_{v \in G} \binom{\deg(v)}{k}$$

- In other words, an answer to the high sensitivity *k*-star query can be derived from results of the degree sequence estimator.

- Would be interesting to compare error of this approach with instance-based noise approach of **[Behoora, PVLDB 11].**

Other work on releasing network statistics

- **[Rastogi, PODS 09]** Subgraph counting queries under an alternative model of *adversarial privacy*. Expected error $\Theta(\log^2 n)$ instead of $\Theta(n)$ for restricted class of adversaries.

- **[Machanavajjhala, PVLDB 11]** Investigates recommender systems that use friends' private data to make recommendations.

  - Lower bound on accuracy of differentially private recommender

  - Experimental analysis shows poor utility under reasonable privacy.

Open questions

- For graph analysis X, how accurately can we estimate X under edge or node differential privacy?
- Lower bounds on accuracy under node 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?

Outline of tutorial

- Privately Managing Enterprise Network Data
- Personal Privacy in Online Social Networks
  - Information sharing in social networks
  - Understanding your privacy risk
  - Managing your privacy control
  - Summary and open questions

Outline of tutorial

- Privately Managing Enterprise Network Data
- Personal Privacy in Online Social Networks
  - Information sharing in social networks
    - What is privacy risk to online social-networking users
    - The sad situation
  - Understanding your privacy risk
  - Managing your privacy control
  - Summary and open questions

Information sharing in social networks

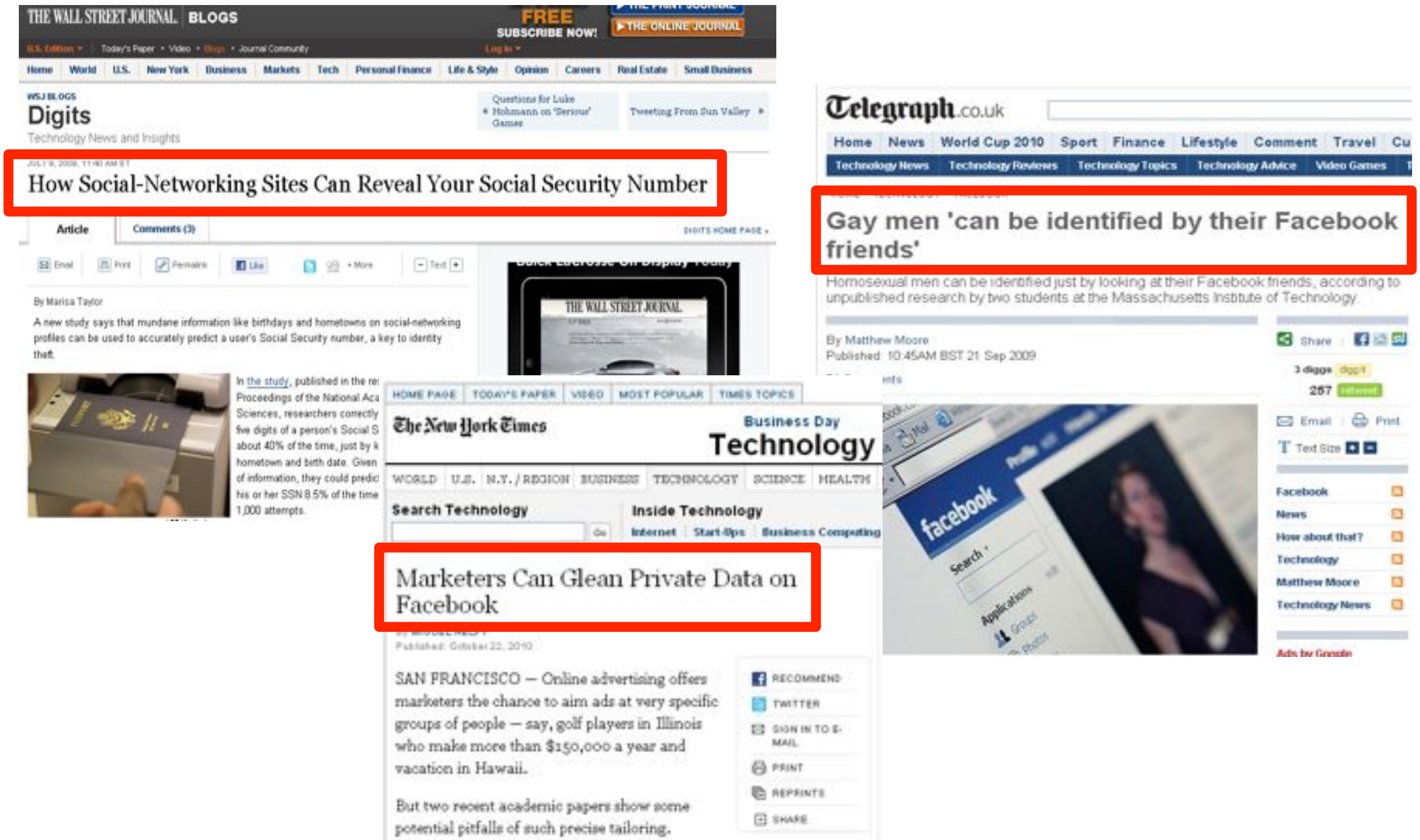
Millions of users share details of their personal lives with vast networks of friends, and often, strangers

Courtesy to: http://www.contrib.andrew.cmu.edu/%7Egct/mygroup.html

What is privacy risk to social-networking users?

The information you share explicitly, e.g., name, age, gender, phone, address, employer, etc. can lead to identity theft.

What is privacy risk to social-networking users?



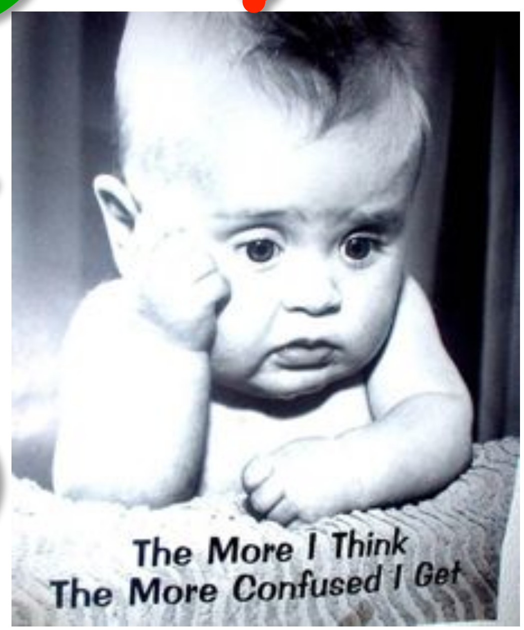
The information you did not share explicitly can also be derived from your public profile, friendship connections or even micro-targeted advertising systems.

The sad situation…

My God! What information I have shared all these years and who can view these information?

All my friends have shared their hometown and phone number, maybe I should also do this?

I enjoyed sharing my daily activities with the World! But any adverse effects?



How to prevent my ex from seeing my status updates?

How to hide my friend list in the search results?

How to prevent the applications my friends installed from accessing my information?

The sad situation… (cont.)

- You have control on what information you want to share, who you want to connect with

- You do not have comprehensive and accurate idea of the information you have explicitly and implicitly disclosed

- Setting online privacy is time consuming and many of you choose to accept the default setting

- Eventually you lose control….and you are facing privacy risk

Outline of tutorial

- Privately Managing Enterprise Network Data
- Personal Privacy in Online Social Networks
  – Information sharing in social networks
  – Understanding your privacy risk
    - Privacy risk due to what you shared explicitly
    - Privacy risk due to what you shared implicitly
    - Tools to visualize your privacy policies
  – Managing your privacy control
  – Summary and open questions

Privacy risk due to what you shared explicitly

- Privacy risk is measured by Privacy Score [Liu, ICDM 09]
- Privacy score takes into account what info you've shared and who can view that info

Basic premises of privacy score

- **Sensitivity:** The more sensitive the information revealed by a user, the higher his privacy risk.

mother's maiden name is more sensitive than ***mobile-phone number***

- **Visibility:** The wider the information about a user spreads, the higher his privacy risk.

home address known ***by everyone*** poses higher risks than ***by friends***

[Liu, ICDM 09]

The framework

name, or gender, birthday, address, phone number, degree, job, etc.

Privacy Score of User *j* due to Profile Item *i*

$$\mathrm{PR}(i, j) = \beta_i \times V(i, j).$$

sensitivity of profile item *i*

visibility of profile item *i*

The framework (cont.)

name, or gender, birthday, address, phone number, degree, job, etc.

Privacy Score of User *j* due to Profile Item *i*

$$\mathrm{PR}(i,j) = \beta_i \times V(i,j).$$

sensitivity of profile item *i*

visibility of profile item *i*

Overall Privacy Score of User *j*

$$\mathrm{PR}(j) = \sum_i \mathrm{PR}(i,j) = \sum_i \beta_i \times V(i,j).$$

The item response theory (IRT) approach

	User_*1*					User_*j*				User_*N*
Profile Item_*1* *(birthday)*	R(*1, 1*)	R(*1, 2*)								R(*1, N*)
Profile Item_*i* *(cell phone #)*						R(*i, j*)				
Profile Item_*n*	R(*n, 1*)									R(*n, N*)

share, R(*i, j*) = 1

not share, R(*i, j*) = 0

$$P_{ij} = \Pr\{R(i,j)=1\} = \frac{1}{1+e^{-\alpha_i(\theta_j-\beta_i)}}$$

Profile item's discrimination

User's attitude,
e.g., conservative or extrovert

Profile item's sensitivity

Profile item *i*'s true visibility

Calculating privacy score using IRT

Overall Privacy Score of User *j*

$$\mathrm{PR}(j) = \sum_i \beta_i \times V(i,j)$$

Sensitivity: β_i

Visibility: $V(i,j) = \Pr\{R(i,j)=1\}$

$$P_{ij} = \Pr\{R(i,j)=1\} = \frac{1}{1+e^{-\alpha_i(\theta_j-\beta_i)}}$$

byproducts: profile item's discrimination and user's attitude

All parameters can be estimated using
Maximum Likelihood Estimation and Expectation-Maximization.

Interesting results from user study

Survey

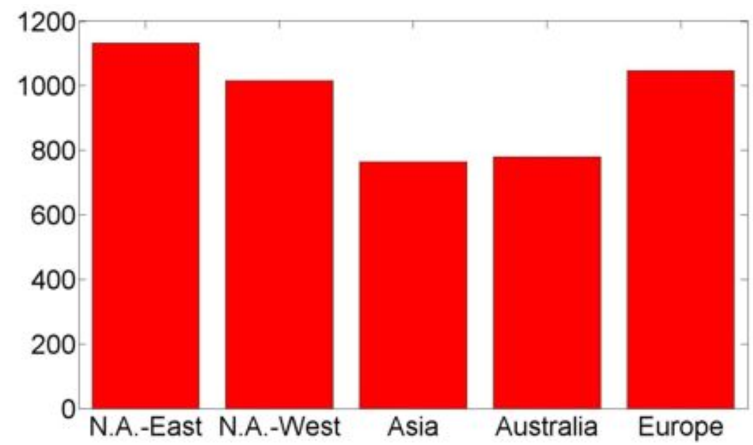
Information-sharing preferences of **153 users** on **49 profile items** such as *name, gender, birthday, political views, address, phone number, degree, job*, etc. are collected.

Statistics

- 49 profile items
- 153 users from 18 countries/regions
- 53.3% are male and 46.7% are female
- 75.4% are in the age of 23 to 39
- 91.6% hold a college degree or higher
- 76.0% spend 4+ hours online per day

Sensitivity of The Profile Items Computed by IRT Model

Average Privacy Scores Grouped by Geo Regions



Outline of tutorial

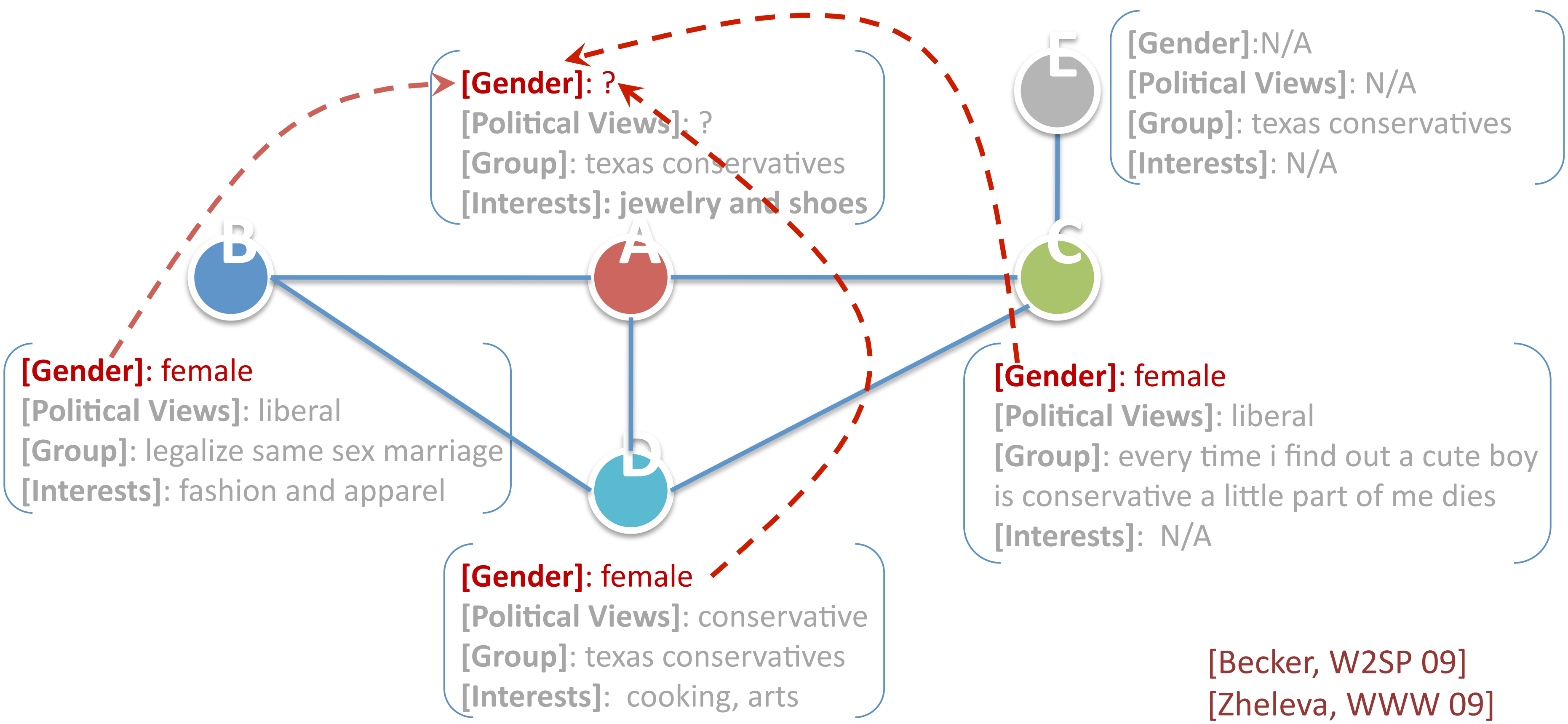
- Privately Managing Enterprise Network Data
- Personal Privacy in Online Social Networks
  - Information sharing in social networks
  - Understanding your privacy risk
    - Privacy risk due to what you shared explicitly
    - Privacy risk due to what you shared implicitly
    - Tools to visualize your privacy policies
  - Managing your privacy control
  - Summary and open questions

Privacy risk due to what you shared implicitly

- Privacy risk is measured by how much your private information can be inferred
- Private information can be inferred from
  - Your public profile, friendships, group memberships, etc.
- Private information can be inferred using
  - Majority voting [Becker, W2SP 09], [Zheleva, WWW 09]
  - Community detection [Mislove, WSDM 10]
  - Classification [Zheleva, WWW 09], [Lindamood, WWW 09]

Inference attack: majority voting

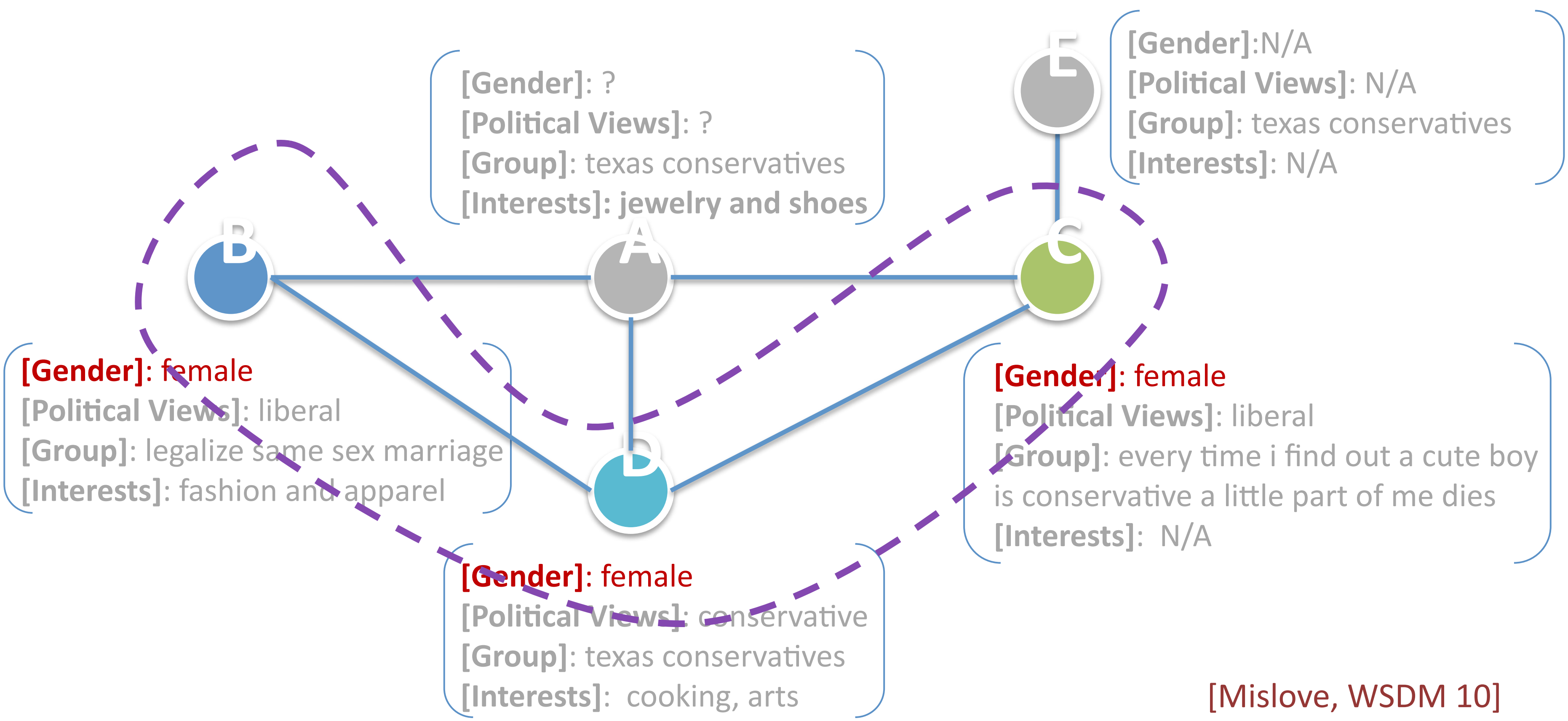
Basic Premise: birds of a feather flock together



[Becker, W2SP 09]
[Zheleva, WWW 09]

Inference attack: community detection

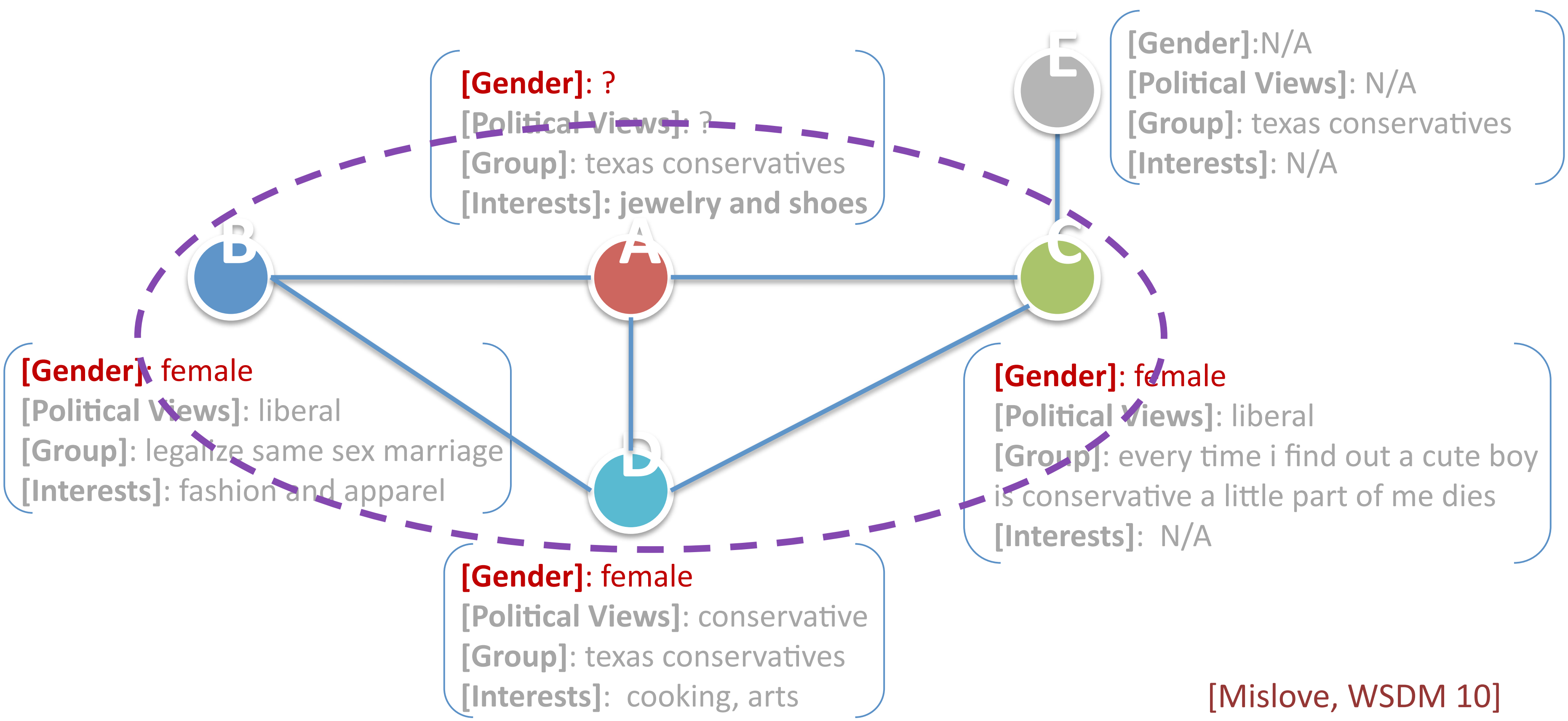
Users with common attributes often form dense communities.



[Mislove, WSDM 10]

Inference attack: community detection

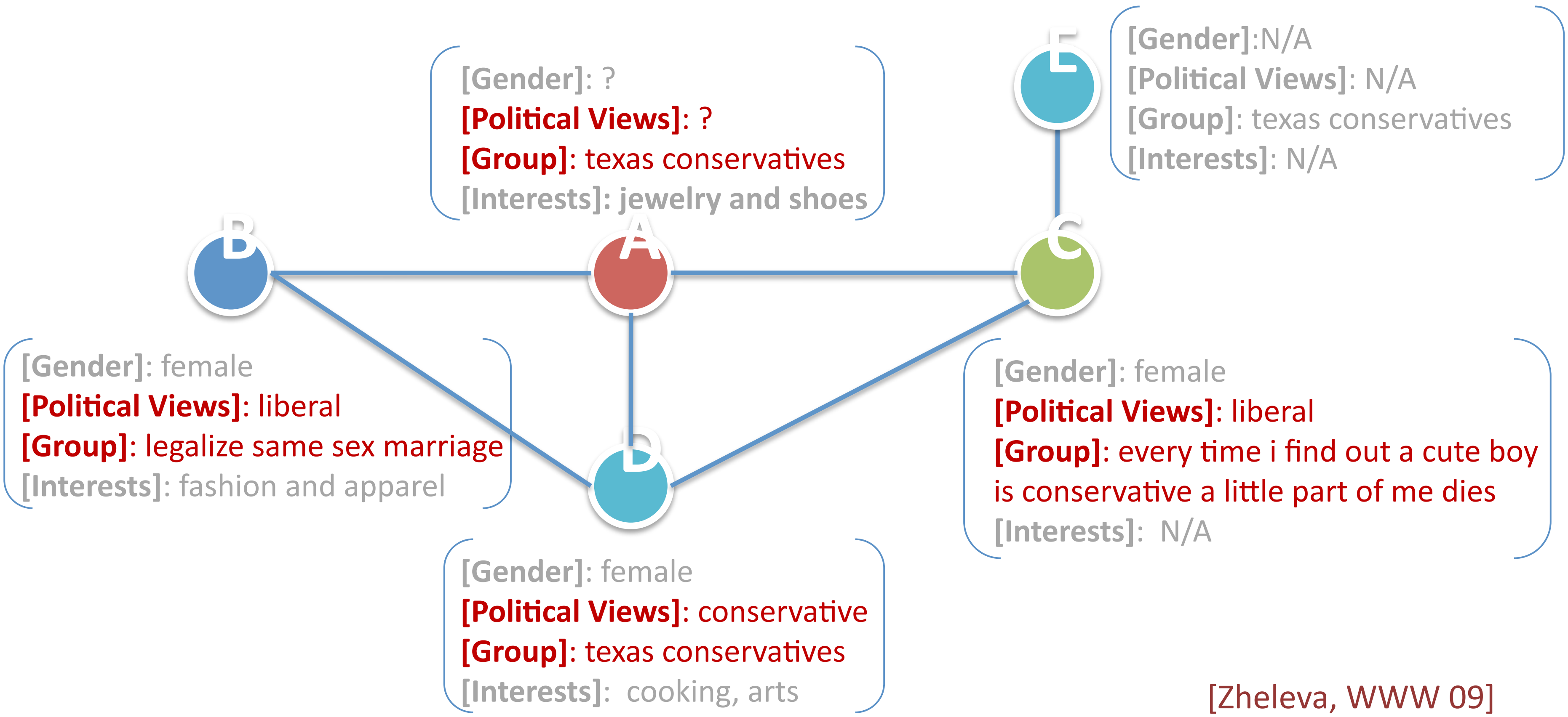
Users with common attributes often form dense communities.



[Mislove, WSDM 10]

Inference attack: classification

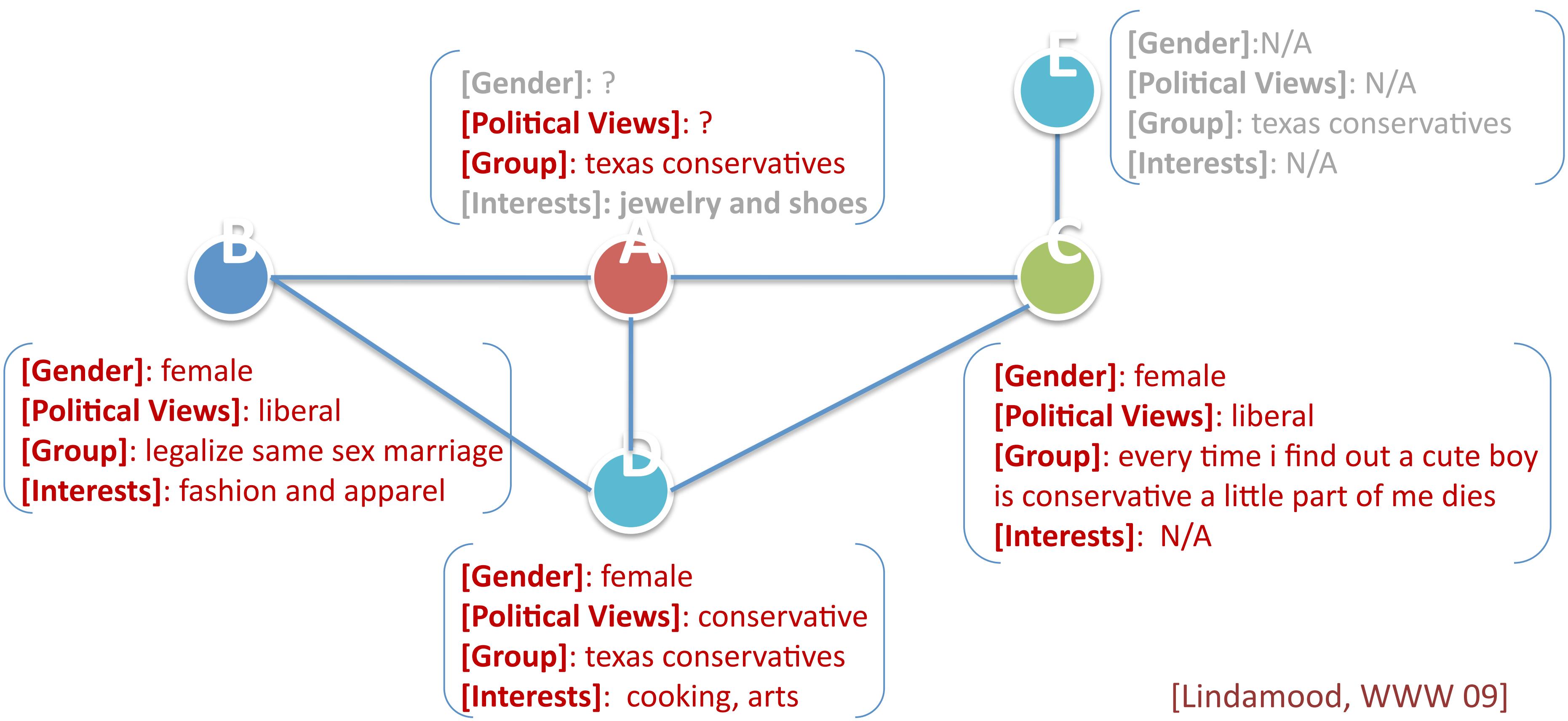
User	legalize same sex marriage	every time i find out a cute boy ...	Texas conservatives	Political views
A	0	0	1	?
B	1	0	0	liberal
C	0	1	0	liberal
D	0	0	1	conservative



[Zheleva, WWW 09]

Inference attack: classification

$$Pr(\text{political views} = \text{'conservative'} \mid \text{group} = \text{'texas conservatives'}, edge_{AB}, edge_{AC}, edge_{AD})$$



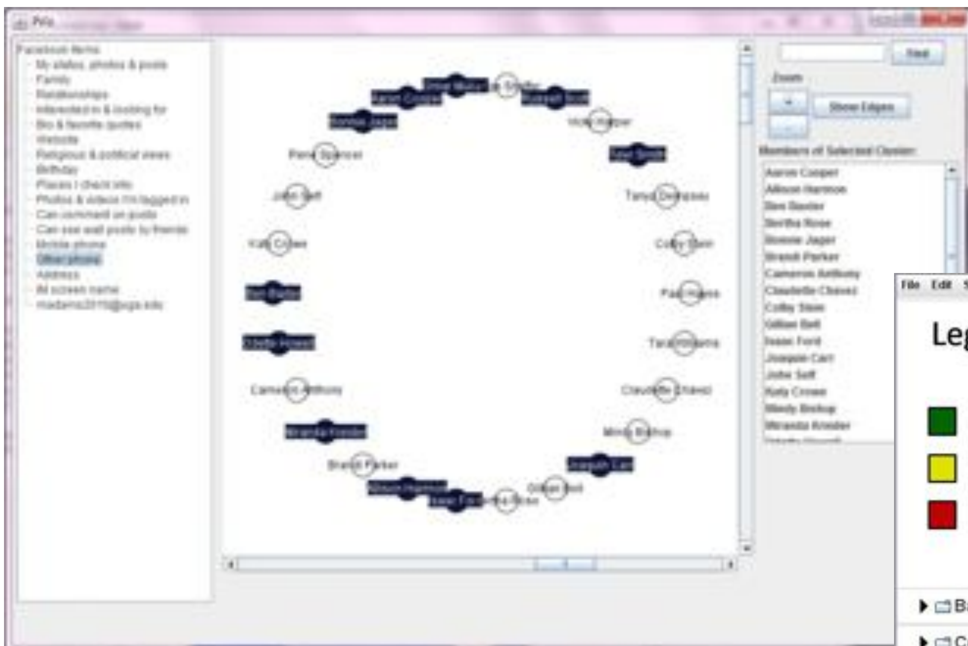
[Lindamood, WWW 09]

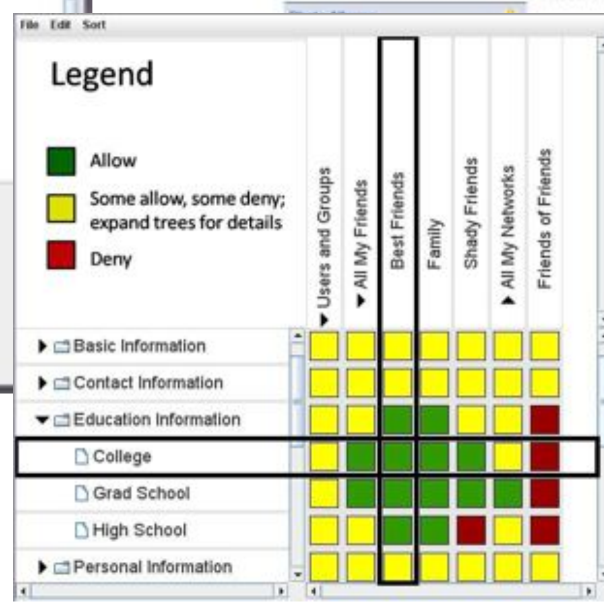
Outline of tutorial

- Privately Managing Enterprise Network Data
- Personal Privacy in Online Social Networks
  - Information sharing in social networks
  - Understanding your privacy risk
    - Privacy risk due to what you shared explicitly
    - Privacy risk due to what you shared implicitly
    - Tools to visualize your privacy policies
  - Managing your privacy control
  - Summary and open questions

Tools to visualize privacy policies

- Visualizations significantly impact users' understanding of their privacy settings [Mazzia, CHI 11], [Lipford, CHI 10]





Outline of tutorial

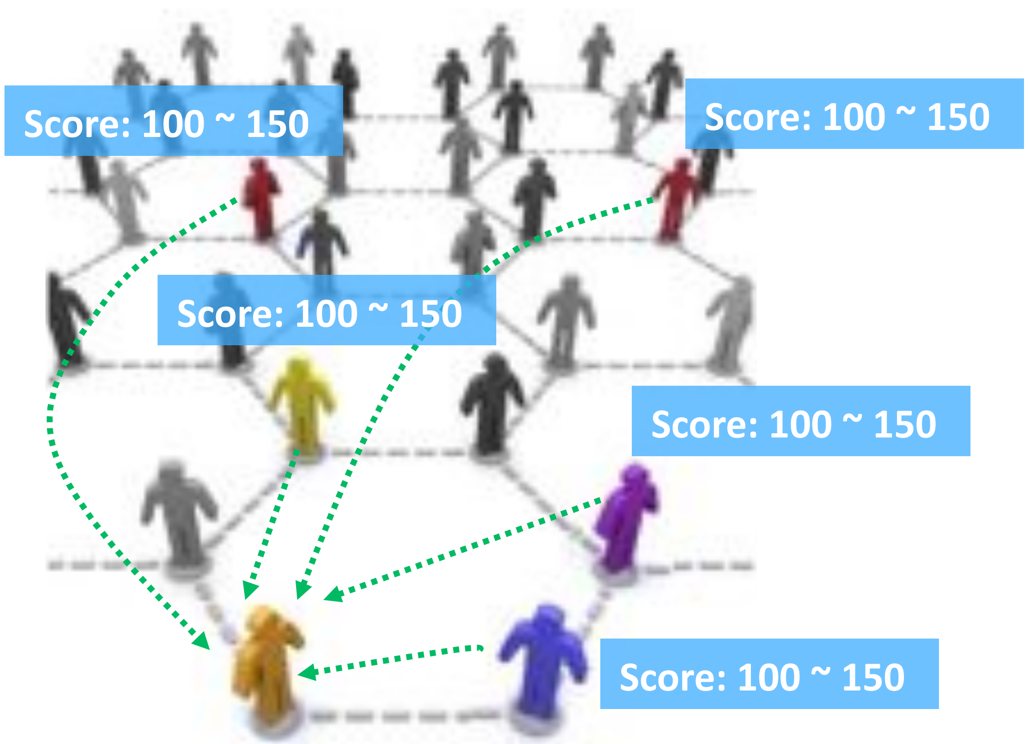
- Privately Managing Enterprise Network Data
- Personal Privacy in Online Social Networks
  - Information sharing in social networks
  - Understanding your privacy risk
  - Managing your privacy control
    - Privacy management for individuals
    - Collaborative privacy management for shared contents
  - Summary and open questions

Privacy management for individuals

- Social navigation [Liu, ICDM 09], [Besmer, SOUPS 10]
- Preventing inference attacks [Lindamood, WWW 09]
- Learning privacy preferences with limited user inputs [Fang, WWW 10], [Shehab, WWW 10]

Social navigation

Social navigation helps users make better privacy decisions using community knowledge and expertise.



[Liu, ICDM 09]

[Besmer, SOUPS 10]

Preventing inference attacks

Remove/hide risky links, profiles or groups that contributed most to the inference attacks.

Pr(political views = 'conservative' | group = ~~'texas conservatives'~~, ~~$edge_{AB}$~~, ~~$edge_{AC}$~~, ~~$edge_{AD}$~~)

[Lindamood, WWW 09]

Learning privacy preferences

Learning privacy preferences with limited user inputs and automatically configure privacy settings for the user.

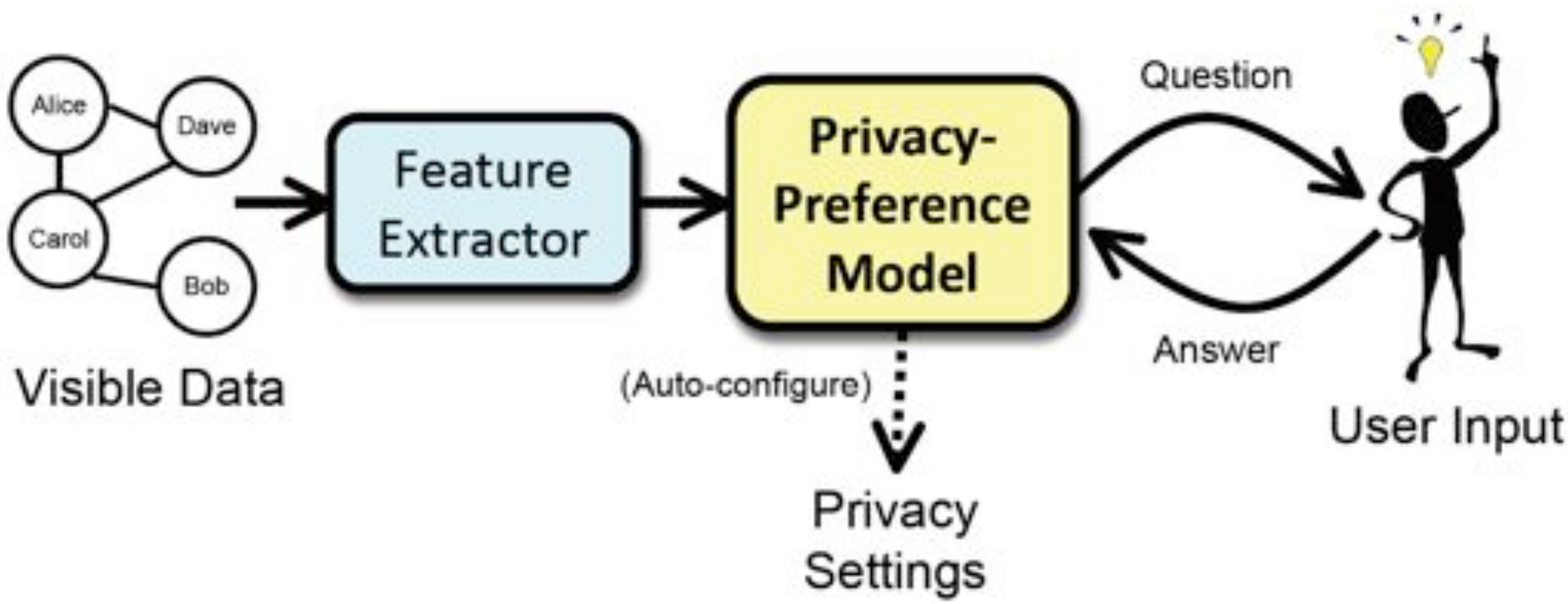


Figure courtesy to Lujun Fang and Kristen LeFevre.

[Fang, WWW 10]

[Shehab, WWW 10]

The framework

- View privacy preference model as a classifier
  - View each friend as a feature vector
  - Predict class label (allow or deny; share or not share)

- Key Design Questions:
  - How to construct features for each friend?
  - How to solicit user inputs in order to get labeled data?

Constructing features for each friend

friends	Age	Sex	G_0	G_1	G_2	G_{20}	G_{21}	G_{22}	G_3	Obama Fan	Pref. Label (*DOB*)
(Alice)	25	F	0	1	0	0	0	0	0	1	**allow**
(Bob)	18	M	0	0	1	1	0	0	0	0	**deny**
(Carol)	30	F	1	0	0	0	0	0	0	0	**?**

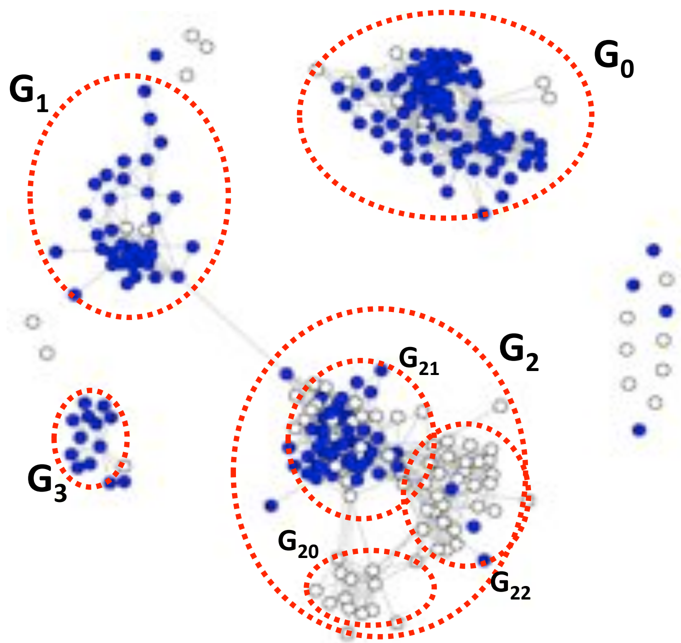


Figure courtesy to Lujun Fang and Kristen LeFevre.

[Fang, WWW 10]
[Shehab, WWW 10]
also see [Jones, SOUPS 10]
also see [Danezis, AISec 09]

Soliciting user inputs

- Ask user to label specific friends
  - e.g., “Would you like to share your Date of Birth with Alice Adams?”

- Choose informative friends using an active learning approach
  - Uncertainty sampling

Uncertainty sampling

- Start with labeled friends F_L and unlabeled friends F_N
- Sampling proceeds in rounds
  - Ask user to label one friend f from F_N
  - f chosen based on uncertainty estimate:
    - Train Bayesian classifier using F_L
    - For each f in F_N, estimate P_{allow}, P_{deny}
    - Choose f in F_N that maximizes
      $Uncertainty = -P_{allow} \log P_{allow} - P_{deny} \log P_{deny}$
- User can quit at any time
- Train preference model (final classifier) using F_L
  - Use to label friends in F_N

Outline of tutorial

- Privately Managing Enterprise Network Data
- Personal Privacy in Online Social Networks
  – Information sharing in social networks
  – Understanding your privacy risk
  – Managing your privacy control
    - Privacy management for individuals
    - Collaborative privacy management for shared contents
  – Summary and open questions

Collaborative privacy management

Photos (or other shared content) uploaded to social networking sites are usually controlled by single users who are not the actual or sole stakeholders.

Collaborative privacy management (cont.)

- The Challenge
  - Each co-owner might have a different and possibly contrasting privacy preference
  - How to choose privacy setting to maximize overall benefits?
- An attempt: clarke tax mechanism [Squicciarini, WWW 09]
  - each owner indicates her perceived benefit at each privacy level (*share with no one, share with friends, etc.*)
  - the system finds the best privacy preference that maximizes the overall social benefit
  - each owner pays certain tax to the system to compensate others' lose
  - the mechanism prevents an owner from untruthfully declaring her benefit to manipulate outcomes at her advantage

[Squicciarini, WWW 09]

Outline of tutorial

- Privately Managing Enterprise Network Data
- Personal Privacy in Online Social Networks
  - Information sharing in social networks
  - Understanding your privacy risk
  - Managing your privacy control
    - Privacy management for individuals
    - Collaborative privacy management for shared contents
  - Summary and open questions

Summary

- You have certain control of the info you are sharing
- You often cannot estimate the long term risk vs. short term gain
- Algorithms to measure potential privacy risks due to info shared either explicitly or implicitly
- Models to alleviate your burden on privacy management

Open questions

- A widely accepted privacy score that boosts public awareness of the privacy risk

- An end-to-end **practical** system to measure and manage privacy online

Privately managing enterprise network data	Personal Privacy in Online Social Networks
Data: Enterprise collects data or observes interactions of individuals.	**Data:** Individuals contribute their data thru participation in OSN.
Control: Enterprise controls dissemination of information.	**Control:** Individuals control their connections, interactions, visibility.
Goal: permit analysis of aggregate properties; protect facts about individuals.	**Goal:** reliable and transparent sharing of information.
Challenges: privacy for networked data, complex utility goals.	**Challenges:** system complexity, leaks thru inference, unskilled users.

Open questions and future directions

- Anonymity: models of adversary knowledge, new attacks, new network transformations, improved utility evaluation.

- Differential privacy: adapting privacy definition to networks, mechanisms for accurate estimates of new network statistics, synthetic network generation, error-optimal mechanisms,

- Extended data model: attributes on nodes/edges, dynamic network data.

References

- [Backstrom, WWW 07] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In WWW 2007.
- [Becker, W2SP 09] J. Becker and H. Chen. Measuring privacy risk in online social networks. In W2SP 2009.
- [Behoora, PVLDB 11] I. Behoora, V. Karwa, S. Raskhodnikova, A. Smith, G. Yaroslavtsev. Private Analysis of Graph Structure. In PVLDB 2011.
- [Besmer, CHI 10] A. Besmer and H. Lipford. Moving beyond untagging: photo privacy in a tagged world. In CHI 2010.
- [Besmer, SOUPS 10] A. Besmer, J. Watson, and H. Lipford. The impact of social navigation on privacy policy configuration. In SOUPS 2010.
- [Bhagat, WWW 10] S. Bhagat, G. Cormode, B. Krishnamurthy, D. Srivastava. Privacy in dynamic social networks. In WWW 2010.
- [Campan, PinKDD 08] A. Campan and T. M. Truta. A clustering approach for data and structural anonymity in social networks. In PinKDD 2008.

References (continued)

- [Chen, Foundations and Trends in Database 09] B. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala. Privacy-Preserving Data Publishing. In Foundations and Trends in Databases 2009.
- [Cheng, SIGMOD 10] J. Cheng, A. Wai-Chee Fu, and J. Liu. K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks. In SIGMOD 2010.
- [Cormode, PVLDB 08] G. Cormode, D. Srivastava, T. Yu, and Q. Zhang: Anonymizing bipartite graph data using safe groupings. In PVLDB 2008.
- [Cormode, PVLDB 09] G. Cormode and D. Srivastava and S. Bhagat and B. Krishnamurthy. Class-based graph anonymization for social network data. In PVLDB 2009.
- [Danezis, AISec 09] G. Danezis. Inferring privacy policies for social networking services. In AISec 2009.
- [Dwork, CACM 10] C. Dwork. A firm foundation for privacy. In CACM 2010.
- [Dwork, STOC 09] C. Dwork and J. Lei. Differential privacy and robust statistics. In STOC 2009.

References (continued)

- [Dwork, TCC 06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In TCC 2006.
- [Fang, WWW 10] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In WWW 2010.
- [Goldenberg, Foundations 10] A. Goldenberg, S. Fienberg, A. Zheng, E. Airoldi. A Survey of Statistical Network Models. In Foundations 2009.
- [Hay, ICDM 09] M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In ICDM 2009.
- [Hay, Privacy-Aware Knowledge Discovery 10] M. Hay and G. Miklau and D. Jensen. Enabling Accurate Analysis of Private Network Data. In Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques 2010.
- [Hay, PVLDB 08] M. Hay, G. Miklau, D. Jensen, and D. Towsley. Resisting structural identification in anonymized social networks. In PVLDB 2008.
- [Hay, PVLDB 10] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially-private queries through consistency. In PVLDB 2010.

References (continued)

- [Hay, VLDBJ 10] M. Hay and G. Miklau and D. Jensen and D. Towsley and C. Li. In VLDB Journal 2010.
- [Hunter, JASA 08] D. Hunter, S. Goodreau, and M. Handcock. Goodness of fit of social network models. In JASA 2008.
- [Jones, SOUPS 10] S. Jones, E. O'Neill. Feasibility of structural network clustering for group-based privacy control in social networks. In SOUPS 2010.
- [Kifer, SIGMOD 11] D. Kifer and A. Machanavajjhala. No Free Lunch in Data Privacy. In SIGMOD 2011.
- [Krishnamurthy, WWW 09] B. Krishnamurthy, C. Wills. Privacy diffusion on the web: a longitudinal perspective. In WWW 2009.
- [Lindamood, WWW 09] J. Lindamood, R. Heatherly, M. Kantarcioglu, B. Thuraisingham. Inferring private information using social network data. In WWW 2009.
- [Lipford, CHI 10] H. Lipford, J. Watson, M. Whitney, K. Froiland, R. Reeder. Visual vs. compact: a comparison of privacy policy interfaces. In CHI 2010.

References (continued)

- [Liu, ICDM 09] K. Liu and E. Terzi. A framework for computing privacy scores of users in online social networks. In ICDM 2009.
- [Liu, Next Generation Data Mining 08] K. Liu, K. Das, T. Grandison, and H. Kargupta. Privacy-Preserving Data Analysis on Graphs and Social Networks. In Next Generation of Data Mining 2008.
- [Liu, SDM 09] L. Liu and J. Wang and J. Liu and J. Zhang. Privacy Preservation in Social Networks with Sensitive Edge Weights. In SDM 2009.
- [Liu, SIGMOD 08] K. Liu and E. Terzi. Towards identity anonymization on graphs. In SIGMOD 2008.
- [Machanavajjhala, PVLDB 11] A. Machanavajjhala, A. Korolova, and A. Das Sarma. Personalized Social Recommendations -- Accurate or Private? In VLDB 2011
- [Mazzia, CHI 11] A. Mazzia, K. LeFevre, and E. Adar. A tool for privacy comprehension. In CHI 2011.

References (continued)

- [Mislove, WSDM 10] A. Mislove, B. Viswanath, K. Gummadi, P. Druschel. You are who you know: Inferring user profiles in online social networks. In WSDM 2010.
- [Narayanan, OAKL 09] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In Security and Privacy 2009.
- [Nissim, STOC 07] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In STOC 2007.
- [Rastogi, PODS 09] V. Rastogi, M. Hay, G. Miklau, and D. Suciu. Relationship privacy: Output perturbation for queries with joins. In PODS 2009.
- [Seigneur, Trust Management 04] J. Seigneur and C. Damsgaard Jensen. Trading privacy for trust, Trust Management 2004
- [Shehab, WWW 10] M. Shehab, G. Cheek, H. Touati, A. Squicciarini, and P. Cheng. Learning based access control in online social networks. In WWW 2010.
- [Smith, IPAM 10] A. Smith. In IPAM Workshop on Statistical and Learning-Theoretic Challenges in Data Privacy 2010.

References (continued)

- [Squicciarini, WWW 09] A. Squicciarini, M. Shehab, F. Paci. Collective privacy management in social networks. In WWW 2009.
- [Wu, Managing and Mining Graph Data 10] X. Wu, X. Ying, K. Liu, and L. Chen. A Survey of Algorithms for Privacy- Preservation of Graphs and Social Networks. In Managing and Mining Graph Data 2010.
- [Wu, SDM 10] L. Wu and X. Ying and X. Wu. Reconstruction from Randomized Graph via Low Rank Approximation. In SDM 2010.
- [Ying, SDM 08] X. Ying and X. Wu. Randomizing social networks: a spectrum preserving approach. In SDM 2008.
- [Ying, SNA-KDD 09] X. Ying and K. Pan and X. Wu and L. Guo. Comparisons of Randomization and K-degree Anonymization Schemes for Privacy Preserving Social Network Publishing. In PinKDD 2009.
- [Zheleva, PinKDD 07] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. In PinKDD 2007.

References (continued)

- [Zheleva, WWW 09] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In WWW 2009.
- [Zhou, ICDE 08] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In ICDE 2009.
- [Zhou, KIS 10] B. Zhou and J. Pei. k-Anonymity and l-Diversity Approaches for Privacy Preservation in Social Networks against Neighborhood Attacks. In Knowledge and Information Systems: An International Journal 2010.
- [Zhou, SIGKDD 08] B. Zhou and J. Pei and W. Luk. A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data. In SIGKDD 2008.
- [Zou, PVLDB 09] L. Zou, L. Chen, and M. T. A. Ozsu. K-automorphism: A general framework for privacy preserving network publication. In PVLDB 2009.