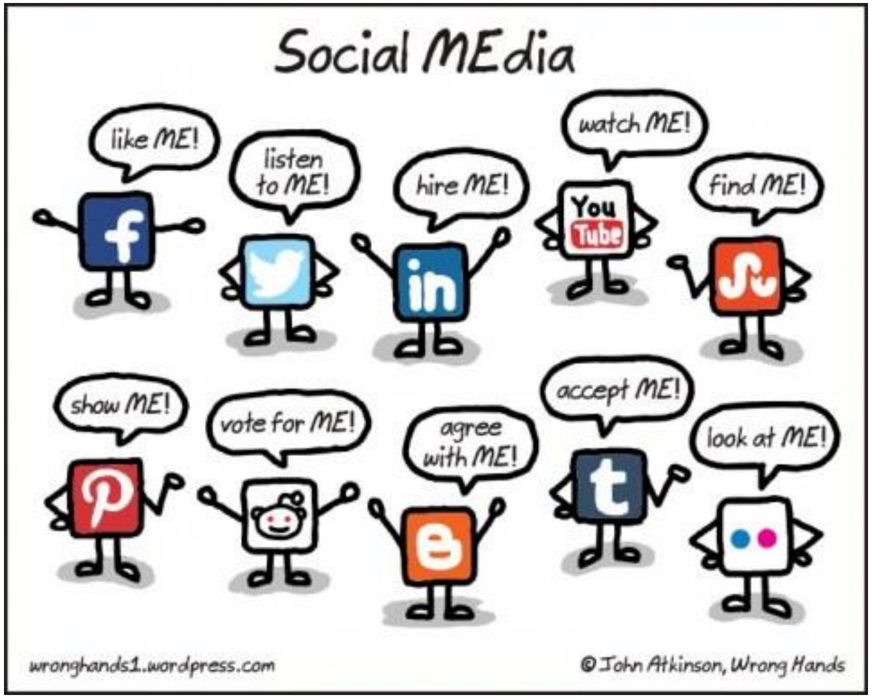


K-Anonymity & Social Networks

CompSci 590.03
Instructor: Ashwin Machanavajjhala

(Some slides adapted from [Hay et al, SIGMOD (tutorial) 2011])

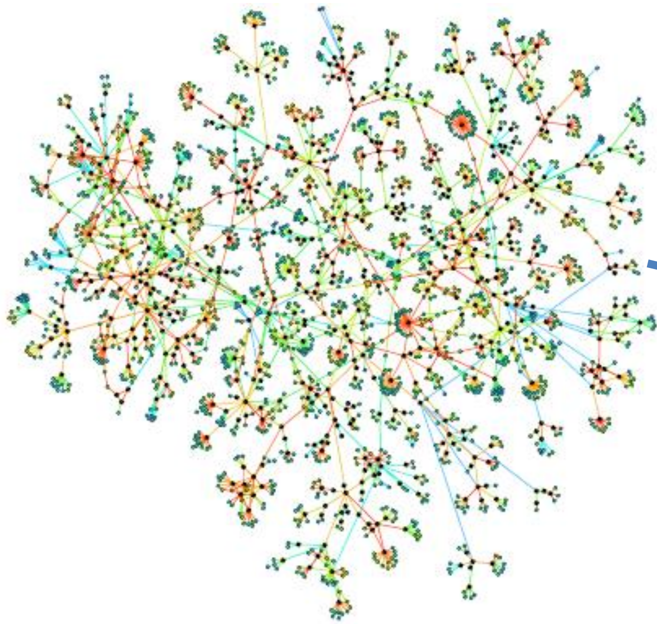
Announcements

- Project ideas are posted on the site.
  - You are welcome to send me (or talk to me about) your own ideas.

http://www.cs.duke.edu/courses/fall12/compsci590.3/project/index.html

Duke UNIVERSITY

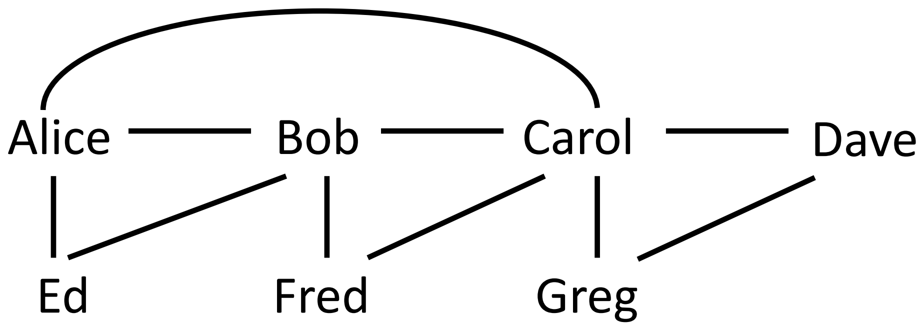
Social Networks are ubiquitous



Mobile communication networks
[J. Onnela et al. PNAS 07]

Sexual & Injection Drug Partners
[Potterat et al. STI 02]

Data Model



Nodes

ID	Age	HIV
Alice	25	+
Bob	19	-
Carol	34	+
Dave	45	+
Ed	32	+
Fred	22	-
Greg	44	-

Edges

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

Why Publish Social Networks?

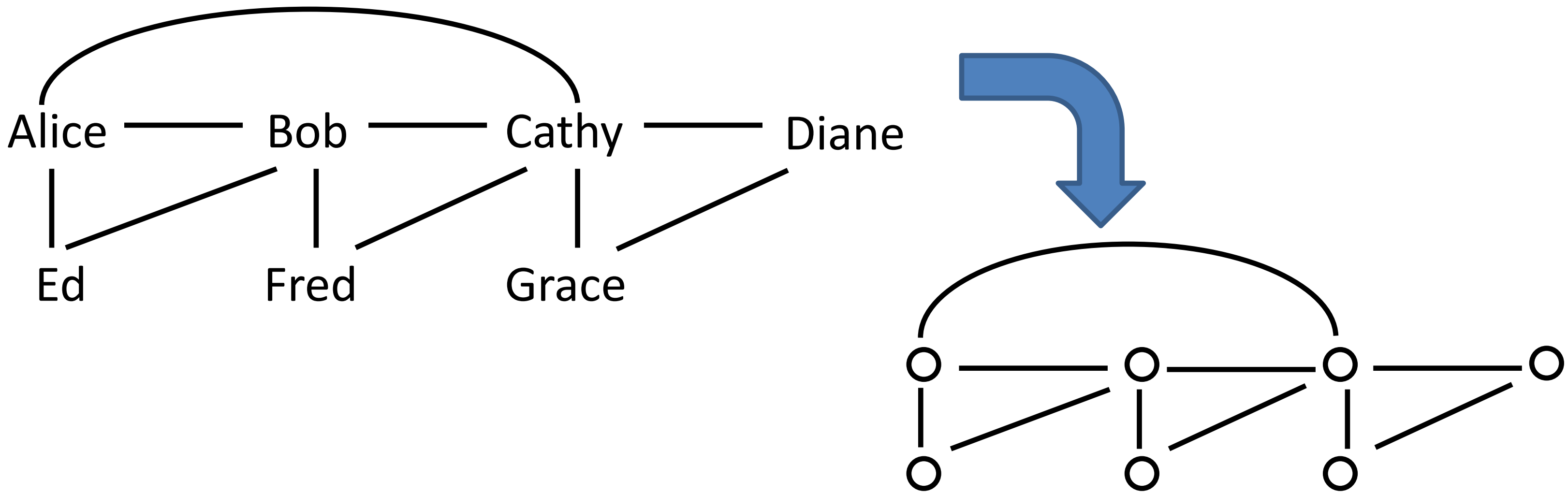
- Statisticians would like to analyze properties of the network

- Example Analyses
  - Degree Distribution
  - Motif analysis
  - Community Structure / Centrality
  - Diffusion on networks
    - Routing, epidemics, information
  - Robustness/ connectivity
  - Homophily
  - Correlation/Causation

What should be protected?

- Node Re-identification: Deduce that node x in the published network corresponds to a real world person Alice.
- Edge Disclosure: Deduce that two individuals Alice and Bob are connected.
- Sensitive property inference: Deduce that Alice is HIV positive.

We already know naïve anonymization does not work!



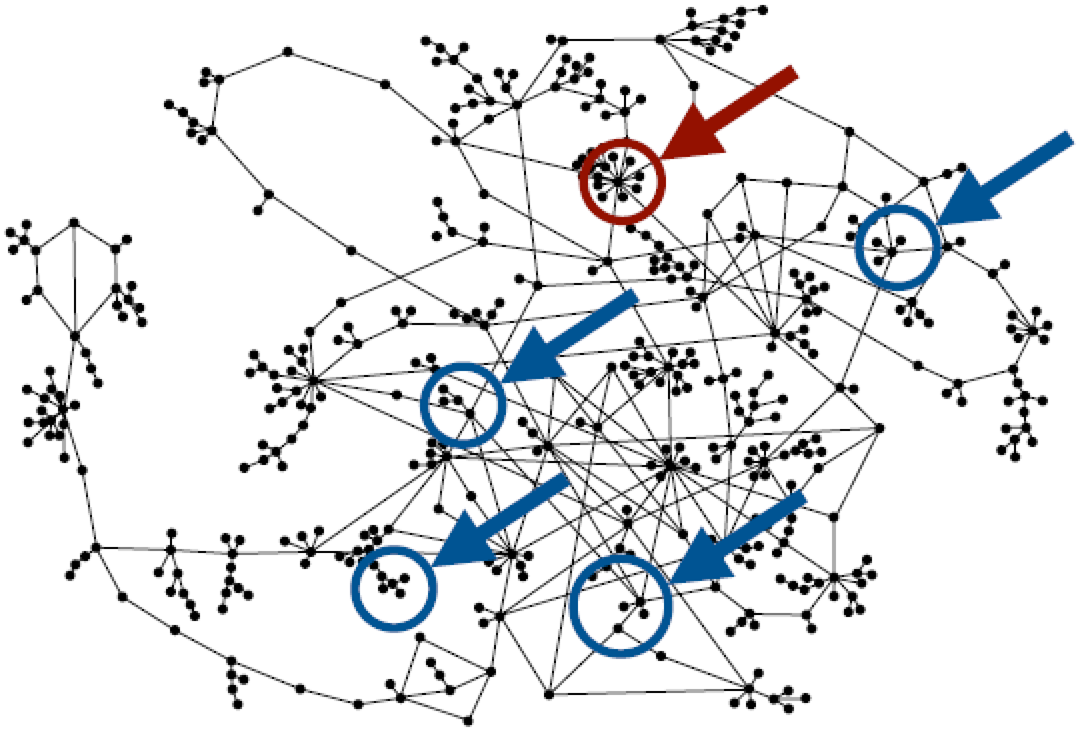
- Naïve Anonymization: replace node identifiers with random numbers.

- Cathy and Alice can identify themselves based on their degree.
- They can together identify Bob and Ed.
- Thus they can deduce Bob and Ed are connected by an edge.

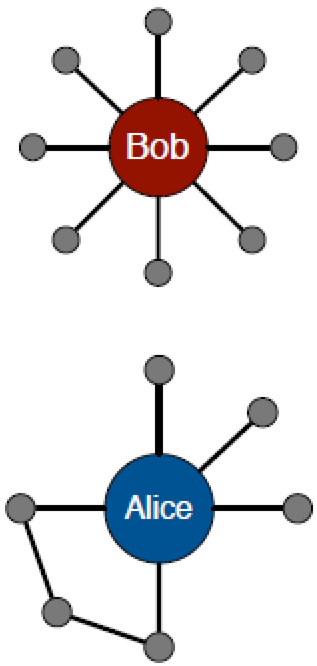
Attacks

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

unique or partial node re-identification



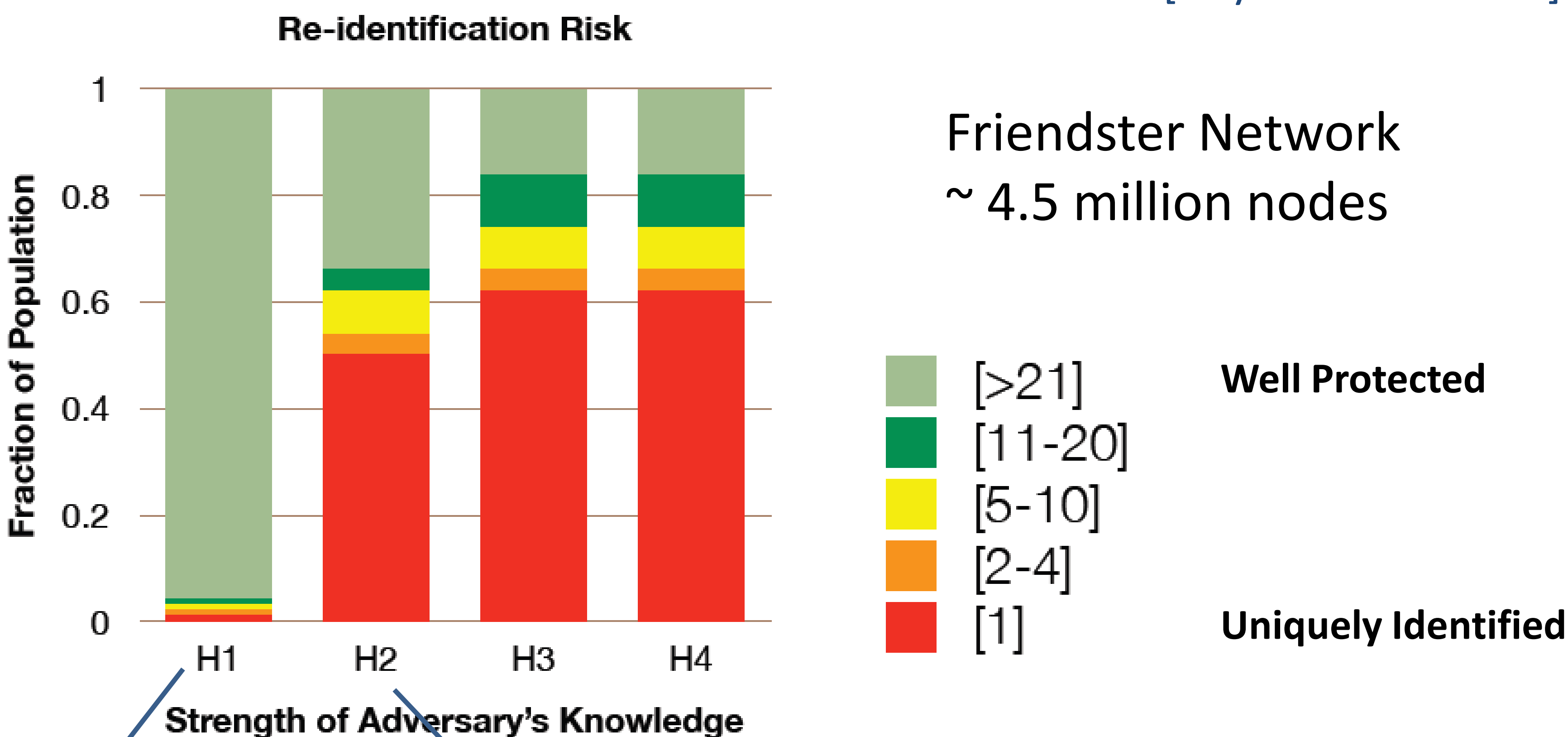
Naively Anonymized Network



External information

Local structure is highly identifying

[Hay et al PVLDB 08]

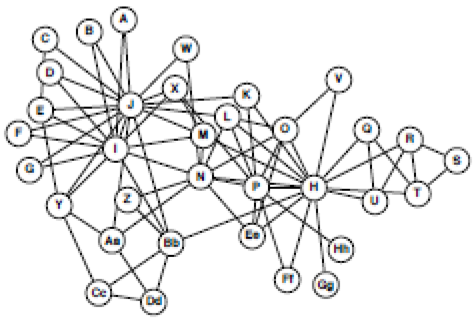


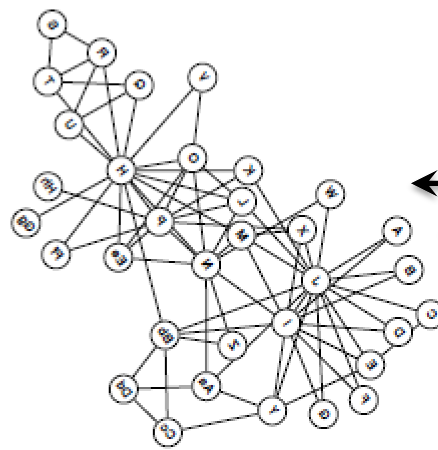
Friendster Network
~ 4.5 million nodes

Node Degree

Neighbor's Degree

Protecting against attacks





Researcher

Transformed Network

- transformations obscure identifying features
- preserve global properties.

Common Problem Formulation

Given input graph G,

- Consider the set of graphs $\mathscr{G}$ such that each G* in $\mathscr{G}$ is reachable from G by certain **graph transformations**.
- Find G* in $\mathscr{G}$ such that it satisfies **anonymity(G*, …)**.
- G* minimizes the **distance(G, G*)**.

Anonymity means …

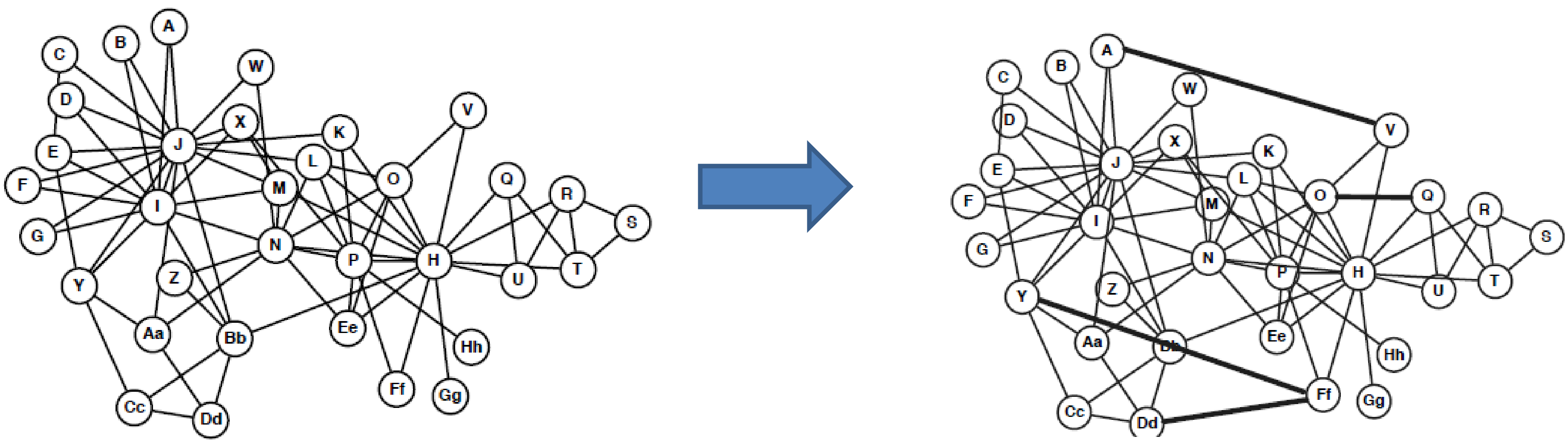
- What do you want to protect ?
  - Node re-identification
  - Edge disclosure

- What can attacker use to break anonymity?
  - attributes
  - Degree
  - Degrees of neighbors
  - Subgraph of neighboring nodes
  - Structural knowledge beyond neighbors.

Distance means …

- No common single measure for utility of the anonymized graph.
- Common approach: empirically compare transformed graph to original graph in terms of various network properties.
  - Degree distribution
  - Path length distribution
  - Clustering coefficient
  - …

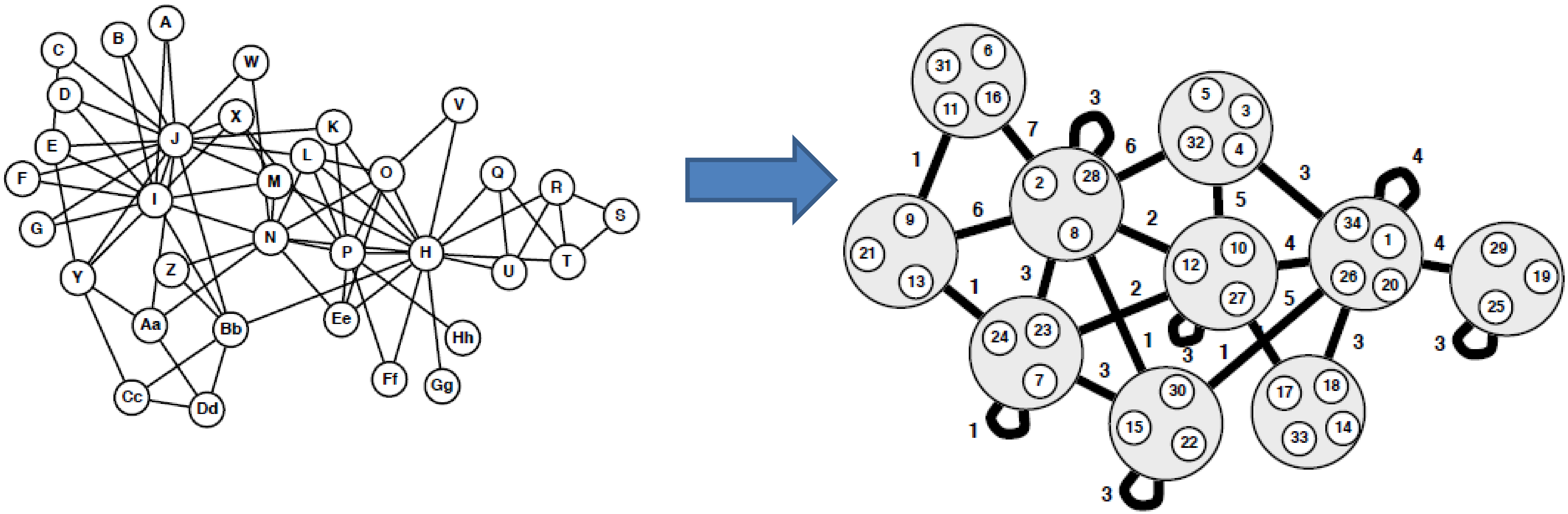
Kinds of Transformations: Directed Alteration



Transform the network by adding or removing edges

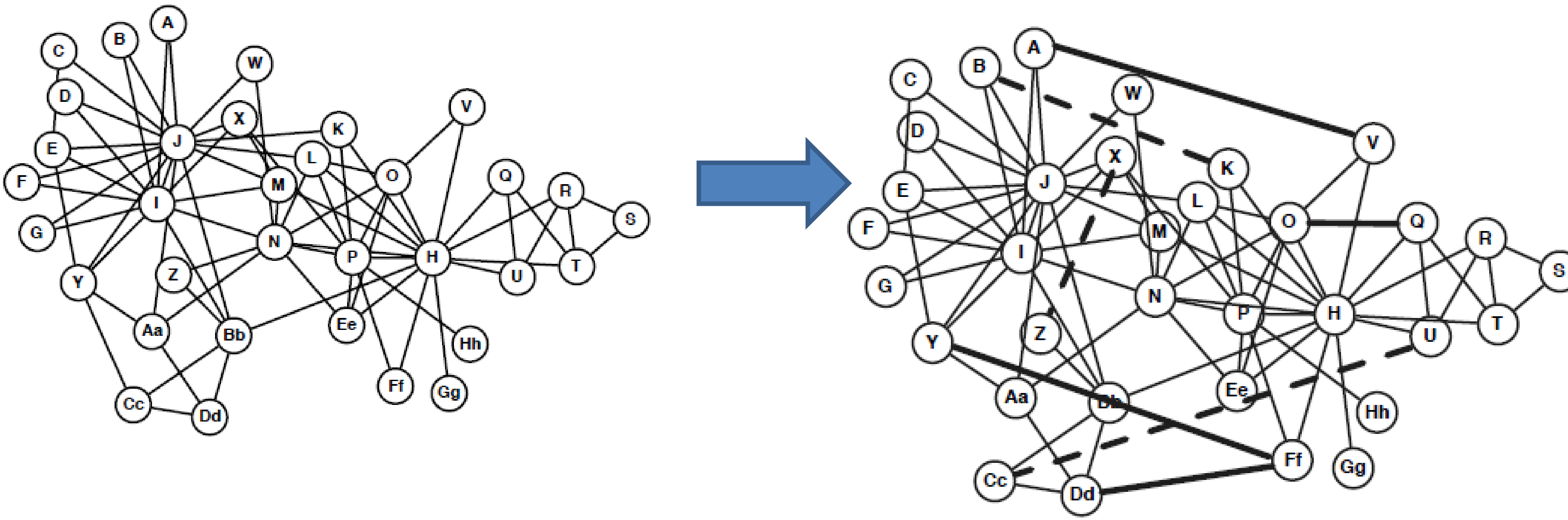
Duke
UNIVERSITY

Kinds of Transformations: Generalization



Transform graph by clustering nodes into groups.

Kinds of Transformations: Randomized Alteration



Transform graph by stochastically adding, removing, or rewiring edges .

Duke UNIVERSITY

	What is protected?	What attacker may know?	Algorithm Strategy
[Liu et al SIGMOD 08]	Node re-identification	Degree of target node	Directed Alteration
[Zhou et al, ICDE 08]	Nodes and labels	Neighborhood of target node (+ labels)	Directed Alteration
[Zou et al PVLDB 09]	Node re-identification	Any structural Property (k-isomorphism)	Directed Alteration
[Cheng et al SIGMOD 10]	Nodes and edges	Any Structural Property (k-automorphism)	Directed Alteration
[Hay et al VLDBJ 10]	Node re-identification	Any Structural Property	Generalization
[Cormode, PVLDB 08]	Edges	Attributes in a bipartite graph	Generalization
[Ying et al SDM 08]	Edges	Unclear	Randomized alteration
[Liu et al SDM 09]	Edges	Unclear	Randomized alteration

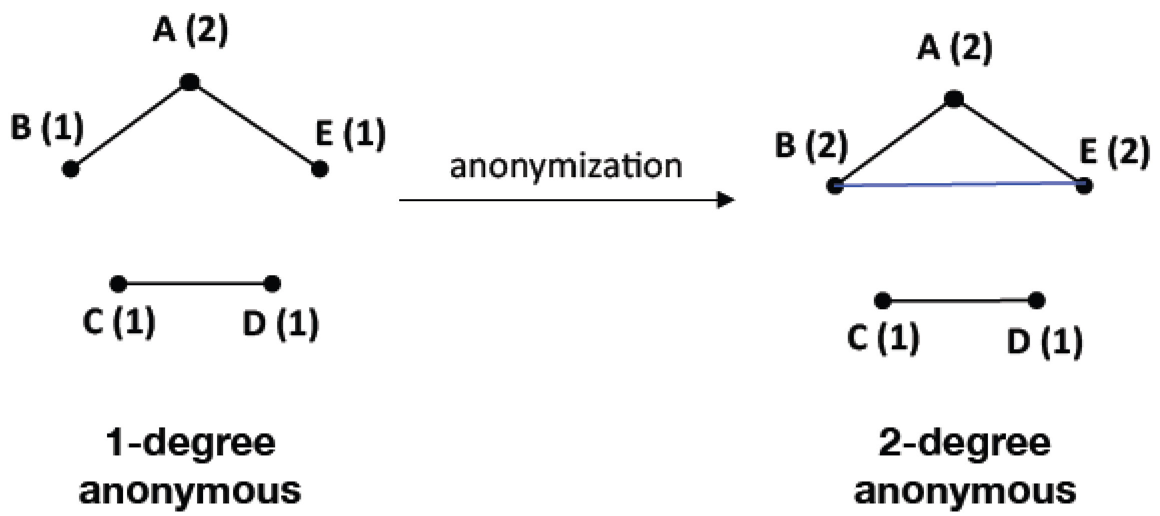
Duke UNIVERSITY

	What is protected?	What attacker may know?	Algorithm Strategy
[Liu et al SIGMOD 08]	**Node re-identification**	**Degree of target node**	**Directed Alteration**
[Zhou et al, ICDE 08]	*Nodes and labels*	*Neighborhood of target node (+ labels)*	*Directed Alteration*
[Zou et al PVLDB 09]	*Node re-identification*	*Any structural Property (k-isomorphism)*	*Directed Alteration*
[Cheng et al SIGMOD 10]	*Nodes and edges*	*Any Structural Property (k-automorphism)*	*Directed Alteration*
[Hay et al VLDBJ 10]	**Node re-identification**	**Any Structural Property**	**Generalization**
[Cormode, PVLDB 08]	Edges	Attributes in a bipartite graph	Generalization
[Ying et al SDM 08]	Edges	Unclear	Randomized alteration
[Liu et al SDM 09]	Edges	Unclear	Randomized alteration

Degree Anonymization

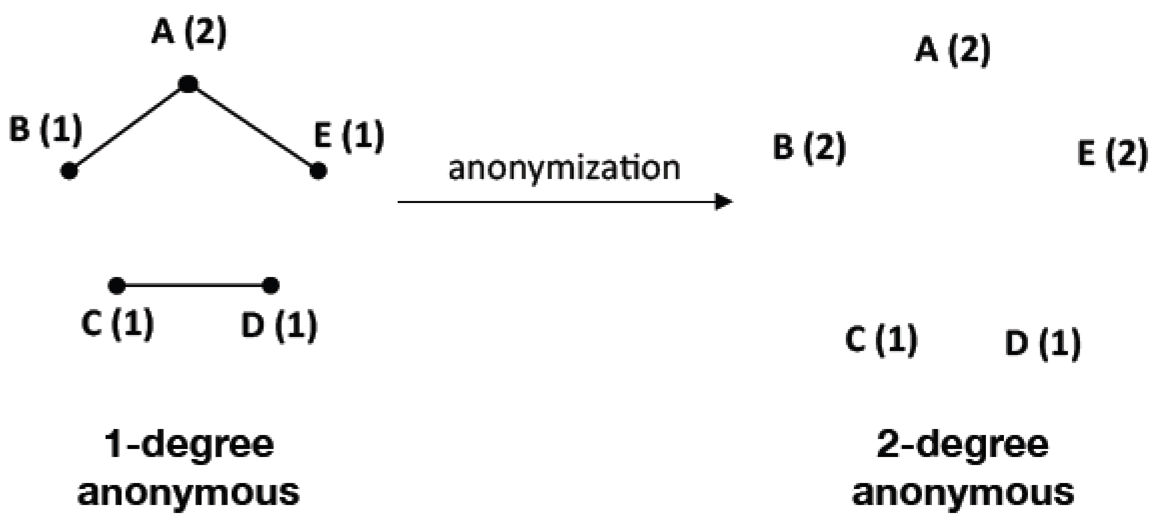
[Liu et al SIGMOD 08]

- Construct a G* such that degree distribution is k-anonymous.



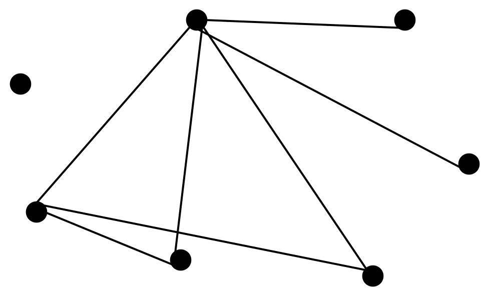
Degree Anonymization

- Step 1: Construct a degree distribution that is close to original distribution, by *minimally increasing* degrees of a few nodes.
- Step 2: Construct a graph satisfying the new degree distribution *close to the original graph* by adding minimum number of edges.



Step 1: k-anonymous degree distribution

minimize $L_1\left(\widehat{\mathbf{d}}-\mathbf{d}\right)=\sum_i\left|\widehat{\mathbf{d}}(i)-\mathbf{d}(i)\right|$



5, 3, 2, 2, 1, 1, 0

- Adding edges means degree only can increase.
- $If\,\widehat{\mathbf{d}}(i)=\widehat{\mathbf{d}}(j),\ with\ i<j,\ then\ \widehat{\mathbf{d}}(i)=\widehat{\mathbf{d}}(i+1)=\ldots=\widehat{\mathbf{d}}(j-1)=\widehat{\mathbf{d}}(j).$

Step 1: k-anonymous degree distribution

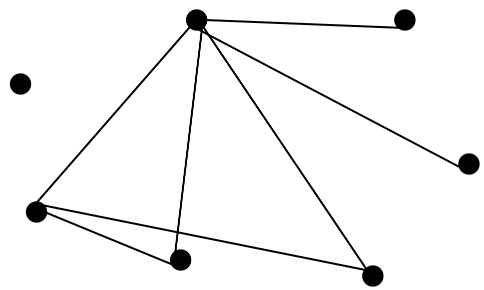
minimize $$L_1\left(\widehat{\mathbf{d}} - \mathbf{d}\right) = \sum_i \left|\widehat{\mathbf{d}}(i) - \mathbf{d}(i)\right|$$

Algorithm?

- Think dynamic programming …

Step 2: Construct a graph with this degree sequence

minimize $L_1\left(\widehat{\mathbf{d}} - \mathbf{d}\right) = \sum_i \left|\widehat{\mathbf{d}}(i) - \mathbf{d}(i)\right|$



5, 3, 2, 2, 1, 1, 0

5, 5, 2, 2, 1, 1, 1

No graph can be realized with this degree sequence

Realizable Degree Sequence

LEMMA 1. *([6]) A degree sequence* $\mathbf{d}$ *with* $\mathbf{d}(1) \geq \ldots \geq \mathbf{d}(n)$ *and* $\sum_i \mathbf{d}(i)$ *even, is realizable if and only if for every* $1 \leq \ell \leq n-1$ *it holds that*

$$\sum_{i=1}^{\ell} \mathbf{d}(i) \leq \ell(\ell-1) + \sum_{i=\ell+1}^{n} \min\{\ell, \mathbf{d}(i)\} \qquad (5)$$

Algorithm ConstructGraph:

- Pick node with the highest degree.
- Add d(v) edges to from v to nodes *w* with the highest degrees.
- Set d(w) = d(w) – 1
- If all degrees are 0 RETURN;
  if some degree is < 0 NOT REALIZABLE

Duke UNIVERSITY

Soundness and Completeness

- Sound: Every graph output by the algorithm satisfies the input degree distribution.
  - Proof ?
- Complete: If there is a graph that satisfies the degree distribution, then the algorithms *does not* output NO.
  - Proof?
  - Think induction …

Step 2: Construct a graph with this degree sequence

Issue 1: Degree sequence may not be realizable.

Issue 2: Realizable degree sequence may not be realizable by only adding edges to original graph G.

(See paper for fixes …)

Protecting against other structural knowledge

[Hay et al VLDBJ10]

- Let G_{naive} be the naïvely anonymized graph.

- Let Q be some structural query
  - $Q_d(x)$ = Degree of the node *x*
  - $Q_{d+}(x)$ = Degrees of neighbors of the node *x*

- $cand_Q(x)$ = set of nodes y in the graph such that Q(x) = Q(y).

Protecting against other structural knowledge

Node anonymity:

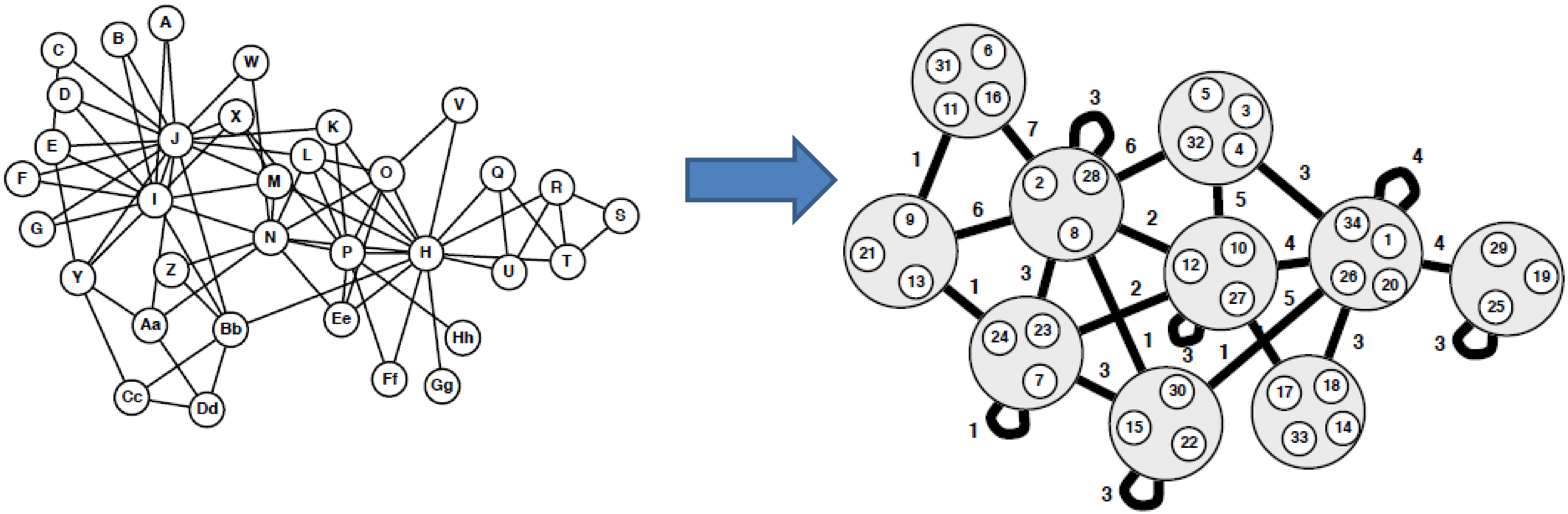
- K-Anonymity: for all x, $|\text{cand}_Q(x)| >= k$

Edge Disclosure: *(more in later classes)*

$$\frac{|\{(u,v) \mid u \in X, v \in Y\}| + |\{(u,v) \mid u,v \in X \cap Y\}|}{|X| \cdot |Y| - |X \cap Y|}$$

where $X = \text{cand}_Q(x)$ and $Y = \text{cand}_Q(y)$.

Ensuring $\text{cand}_Q(x) >= k$



- Each *supernode* has at least k nodes.
- Self loops: number of edges within a super node
- Edges: number of edges between super nodes.

Using a generalized graph

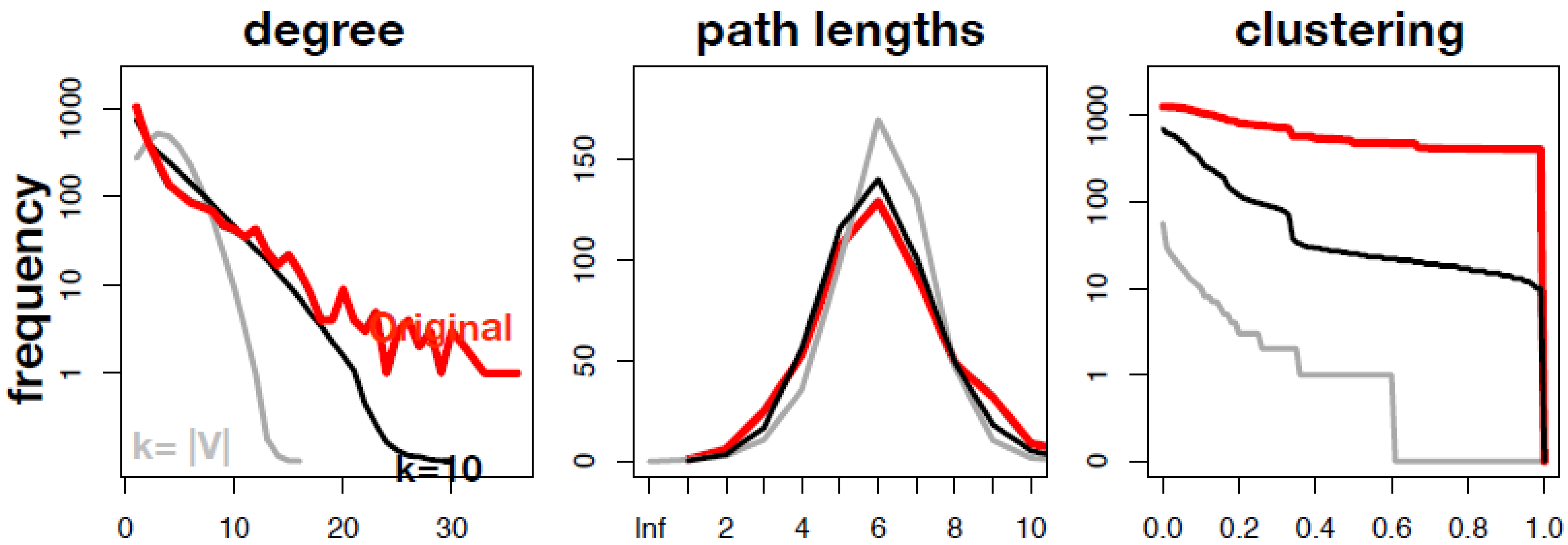
- Many graphs may be generalized to G*

$$|\mathcal{W}(\mathcal{G})| = \prod_{X \in \mathcal{V}} \binom{\frac{1}{2}|X|(|X|-1)}{d(X,X)} \prod_{X,Y \in \mathcal{V}} \binom{|X||Y|}{d(X,Y)}$$

- Run analysis on one or more samples that are consistent with generalized graph.
  - Sample: Pick any graph that are consistent with G* uniformly at random

Duke UNIVERSITY

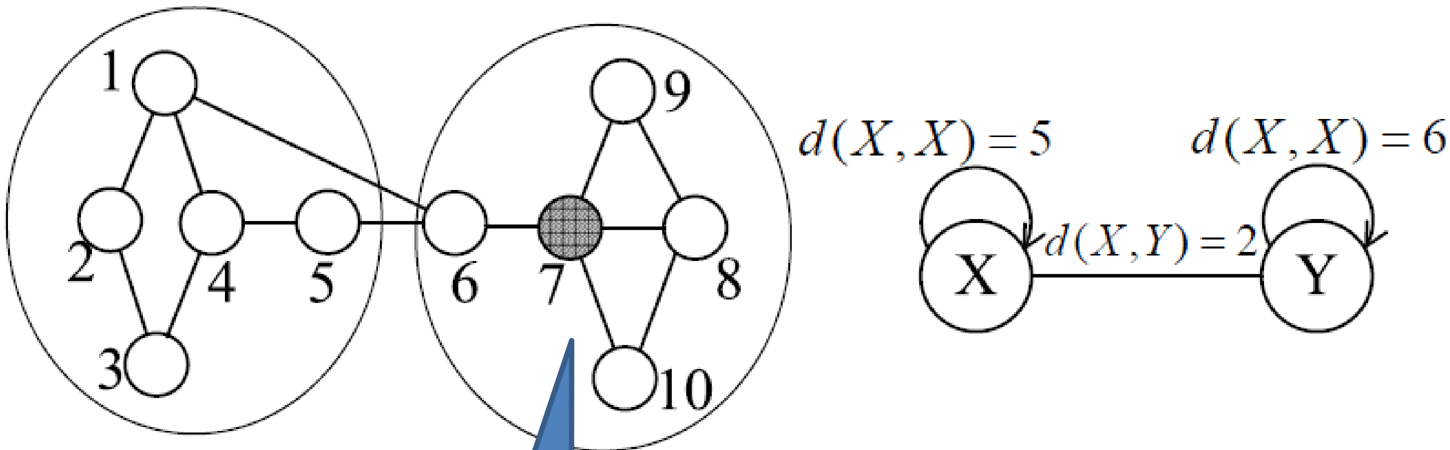
Utility



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

Drawback of Generalization

[Zou et al PVLDB 09]



(a) Naïve Anony zation Network G'

(b) Generalized Network

Lose all the structural information within super node

Duke UNIVERSITY

K-automorphism

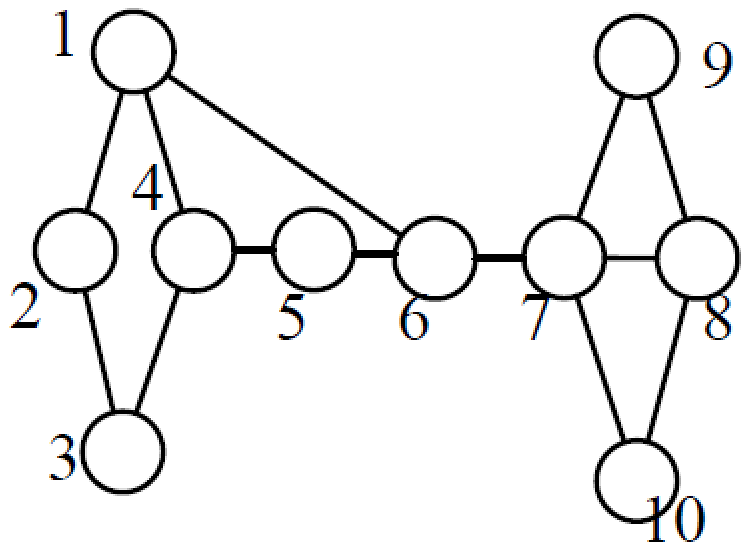
- (non-trivial) Automorphism:
  Given a graph G, there exists f: V → V such that
  (u,v) is an edge in G if and only if (f(u), f(v)) is an edge in G.

- K-Automorphism:
  Given a graph G, there exist K-1 non-trivial automorphisms f_1, f_2, …, f_{k-1} such that for all vertices v, $f_i(v) \neq f_j(v)$

Duke UNIVERSITY

K-automorphism

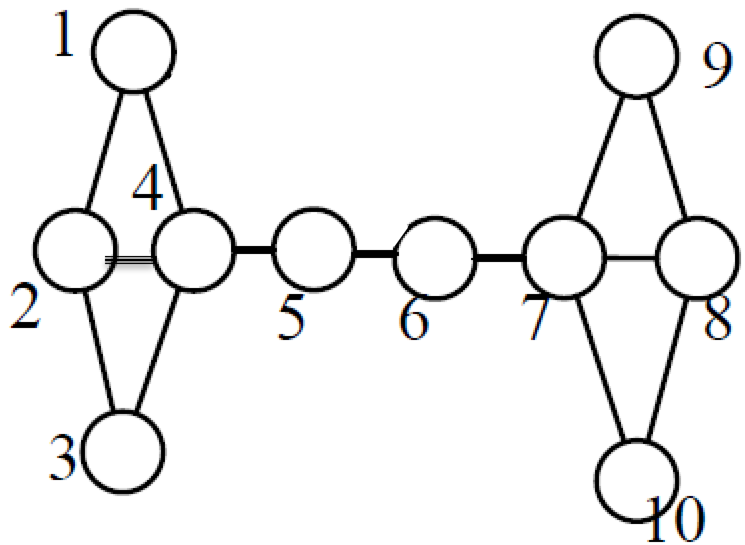
- K-Automorphism:
  Given a graph G, there exist K-1 non-trivial automorphisms f_1, f_2, …, f_{k-1} such that for all vertices v, $f_i(v) \neq f_j(v)$



Not even 2-automorphic

K-automorphism

- K-Automorphism:
  Given a graph G, there exist K automorphisms f1, f2, …, fk such that for all vertices v, $f_i(v) \neq f_j(v)$



This is 2-automorphic

Summary

- Social networks are more susceptible to attacks on anonymity
- Algorithms differ in
  - What is being protected (nodes / edges)
  - What structural property anonymity is based on
  - How the graph is transformed
- But, Anonymity does not guarantee privacy – Next Class.

References

L. Sweeney, "*K-Anonymity: a model for protecting privacy*", IJUFKS 2002

M. Hay, K. Liu, G. Miklau, J. Pei, E. Terzi, "Privacy-Aware Data Management in Information Networks", SIGMOD (tutorial) 2011

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

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

K. Liu & E. Terzi, *"Towards identity anonymization on graphs"*, SIGMOD 2008

M. Hay, G. Miklau, D. Jensen, D. Towsley, & P. Weis. "*Resisting structural re-identification in anonymized social networks.*" PVLDB 2008.

B. Zhou & J. Pei. "*Preserving privacy in social networks against neighborhood attacks*." ICDE 2008.

J. Cheng, A. W. chee Fu, & J. Liu. "*K-isomorphism: privacy preserving network publication against structural attacks*." SIGMOD 2010.

L. Zou, L. Chen, & M. T. Ozsu. "*k-automorphism: a general framework for privacy preserving network publication.*" VLDB, 2009.

Duke UNIVERSITY

References (contd)

L. Liu, J. Wang, J. Liu & J. Zhang. "*Privacy Preservation in Social Networks with Sensitive Edge Weights*". In SDM 2009.

X. Ying and X. Wu. "*Randomizing social networks: a spectrum preserving approach*". In SDM 2008.

G. Cormode, D. Srivastava, T. Yu, and Q. Zhang "*Anonymizing bipartite graph data using safe groupings".* In PVLDB 2008