Anonymization and de-anonymization of graphs

- Reference:
- Towards identity anonymization in social networks (by Kun Liu and Evimaria Terzi, SIGMOD 2008)

BOSTON UNIVERSITY

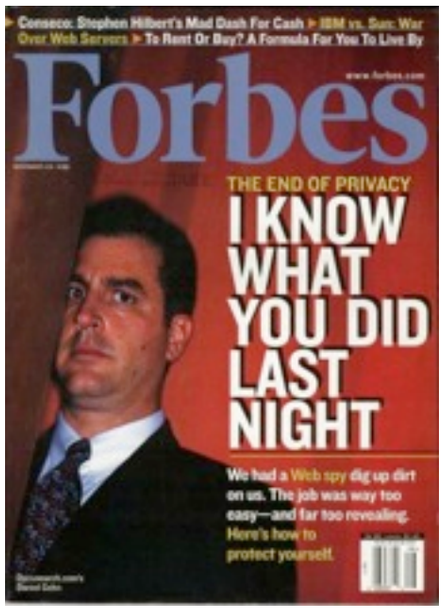
evimaria@cs.bu.edu

Growing Privacy Concerns

- Person specific information is being routinely collected.

"Detailed information on an individual's credit, health, and financial status, on characteristic purchasing patterns, and on other personal preferences is routinely recorded and analyzed by a variety of governmental and commercial organizations."

- M. J. Cronin, "e-Privacy?" Hoover Digest, 2000.



evimaria@cs.bu.edu

Proliferation of Graph Data

http://www.touchgraph.com/

LinkedIn

facebook

myspace.com
a place for friends

Privacy breaches on graph data

- Identity disclosure
  - Identity of individuals associated with nodes is disclosed
- Link disclosure
  - Relationships between individuals are disclosed
- Content disclosure
  - Attribute data associated with a node is disclosed

evimaria@cs.bu.edu

Identity anonymization on graphs

BOSTON UNIVERSITY

evimaria@cs.bu.edu

Identity anonymization on graphs

- Question
  - How to share a network in a manner that permits useful analysis without disclosing the identity of the individuals involved?

BOSTON UNIVERSITY

evimaria@cs.bu.edu

Identity anonymization on graphs

- Question
  - How to share a network in a manner that permits useful analysis without disclosing the identity of the individuals involved?

BOSTON UNIVERSITY

evimaria@cs.bu.edu

Identity anonymization on graphs

- Question
  - How to share a network in a manner that permits useful analysis without disclosing the identity of the individuals involved?

- Observations
  - Simply removing the identifying information of the nodes before publishing the actual graph does not guarantee identity anonymization.

L. Backstrom, C. Dwork, and J. Kleinberg, “Wherefore art thou R3579X?: Anonymized social netwoks, hidden patterns, and structural steganography,” In WWW 2007.

J. Kleinberg, “Challenges in Social Network Data: Processes, Privacy and Paradoxes, ” KDD 2007 Keynote Talk.

evimaria@cs.bu.edu

Identity anonymization on graphs

- Question
  - How to share a network in a manner that permits useful analysis without disclosing the identity of the individuals involved?

- Observations
  - Simply removing the identifying information of the nodes before publishing the actual graph does not guarantee identity anonymization.

L. Backstrom, C. Dwork, and J. Kleinberg, “Wherefore art thou R3579X?: Anonymized social netwoks, hidden patterns, and structural steganography,” In WWW 2007.

J. Kleinberg, “Challenges in Social Network Data: Processes, Privacy and Paradoxes, ” KDD 2007 Keynote Talk.

evimaria@cs.bu.edu

Identity anonymization on graphs

- Question
  - How to share a network in a manner that permits useful analysis without disclosing the identity of the individuals involved?

- Observations
  - Simply removing the identifying information of the nodes before publishing the actual graph does not guarantee identity anonymization.

L. Backstrom, C. Dwork, and J. Kleinberg, “Wherefore art thou R3579X?: Anonymized social netwoks, hidden patterns, and structural steganography,” In WWW 2007.

J. Kleinberg, “Challenges in Social Network Data: Processes, Privacy and Paradoxes, ” KDD 2007 Keynote Talk.

- Can we borrow ideas from ***k***-anonymity?

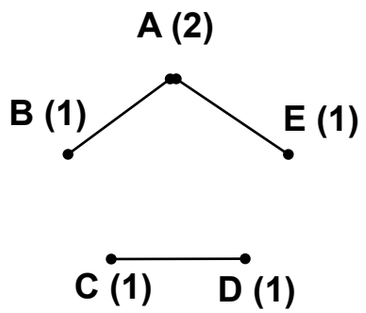
What if you want to prevent the following from happening

- Assume that adversary **A** knows that **B** has 327 connections in a social network!
- If the graph is released by removing the identity of the nodes
  - **A** can find all nodes that have degree 327
  - If there is only one node with degree 327, **A** can identify this node as being **B**.

evimaria@cs.bu.edu

Privacy model

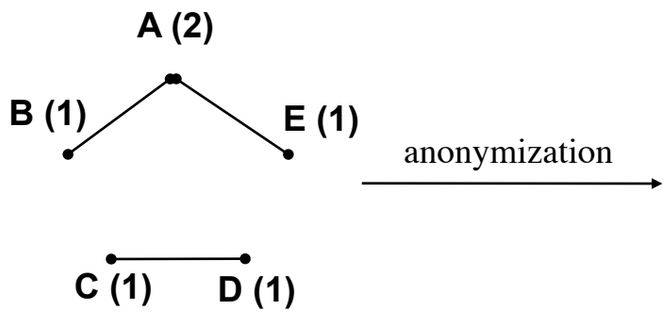
[k-degree anonymity] A graph ***G(V, E)*** is ***k****-degree anonymous* if every node in ***V*** has the same degree as ***k-1*** other nodes in ***V***.



evimaria@cs.bu.edu

Privacy model

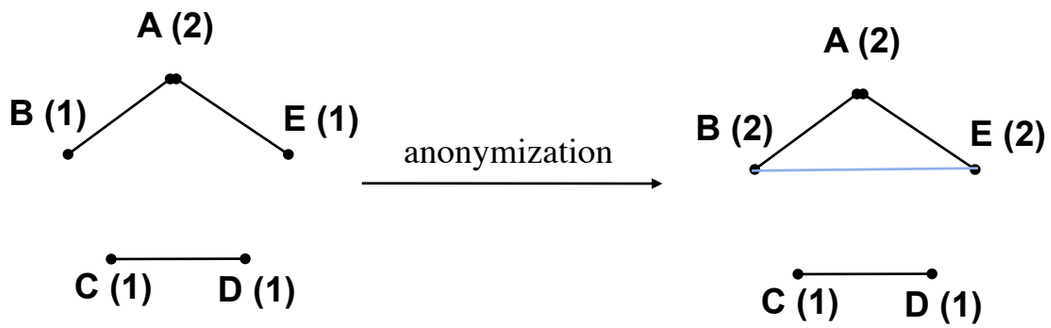
[k-degree anonymity] A graph ***G(V, E)*** is ***k****-degree anonymous* if every node in ***V*** has the same degree as ***k-1*** other nodes in ***V***.



evimaria@cs.bu.edu

Privacy model

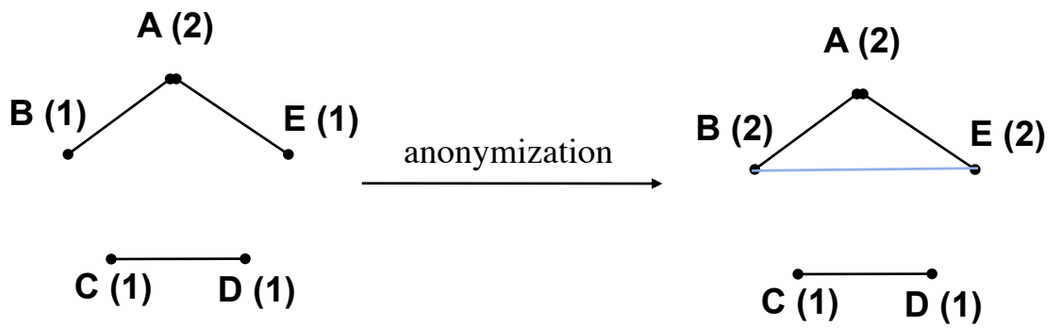
[k-degree anonymity] A graph ***G(V, E)*** is ***k****-degree anonymous* if every node in ***V*** has the same degree as ***k-1*** other nodes in ***V***.



evimaria@cs.bu.edu

Privacy model

[k-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***.



[Properties] It prevents the re-identification of individuals by adversaries with *a priori* knowledge of the degree of certain nodes.

evimaria@cs.bu.edu

Outline

- **Problem definition**
- **Algorithms**
- **Experimental results**

evimaria@cs.bu.edu

Problem Definition

Given a graph ***G(V, E)*** and an integer ***k***, modify ***G*** via a **minimal** set of edge addition or deletion operations to construct a new graph ***G'(V', E')*** such that

1) ***G'*** is ***k***-degree anonymous;

2) ***V' = V***;

3) The ***symmetric difference*** of ***G*** and ***G'*** is as small as possible

- Symmetric difference between graphs ***G(V,E)*** and ***G'(V,E')*** :

$$\mathrm{SymDiff}(G', G) = (E' \setminus E) \cup (E \setminus E')$$

evimaria@cs.bu.edu

Outline

- **Problem definition**
- **Algorithms**
- **Experimental results**

evimaria@cs.bu.edu

GraphAnonymization algorithm

Input: Graph ***G*** with degree sequence ***d***, integer ***k***
Output: ***k***-degree anonymous graph ***G'***

[**Degree Sequence Anonymization**]:

- Contruct an anonymized degree sequence ***d'*** from the original degree sequence ***d***

[**Graph Construction**]:

[**Construct**]: Given degree sequence ***d'***, construct a new graph $G^0(V, E^0)$ such that the degree sequence of G^0 is ***d'***

[**Transform**]: Transform $G^0(V, E^0)$ to $G'(V, E')$ so that ***SymDiff***(G', G) is minimized.

evimaria@cs.bu.edu

Degree-sequence anonymization

[k-anonymous sequence] A sequence of integers ***d*** is ***k-anonymous*** if every distinct element value in ***d*** appears at least ***k*** times.

BOSTON UNIVERSITY

evimaria@cs.bu.edu

Degree-sequence anonymization

[k-anonymous sequence] A sequence of integers ***d*** is ***k-anonymous*** if every distinct element value in ***d*** appears at least ***k*** times.

[100,100, 100, 98, 98,15,15,15]

BOSTON UNIVERSITY

evimaria@cs.bu.edu

Degree-sequence anonymization

[k-anonymous sequence] A sequence of integers ***d*** is ***k-anonymous*** if every distinct element value in ***d*** appears at least ***k*** times.

[100,100, 100, 98, 98,15,15,15]

[degree-sequence anonymization] Given degree sequence ***d***, and integer ***k***, construct ***k***-anonymous sequence ***d'*** such that ***||d'-d||*** **is minimized**

evimaria@cs.bu.edu

Degree-sequence anonymization

[k-anonymous sequence] A sequence of integers ***d*** is ***k-anonymous*** if every distinct element value in ***d*** appears at least ***k*** times.

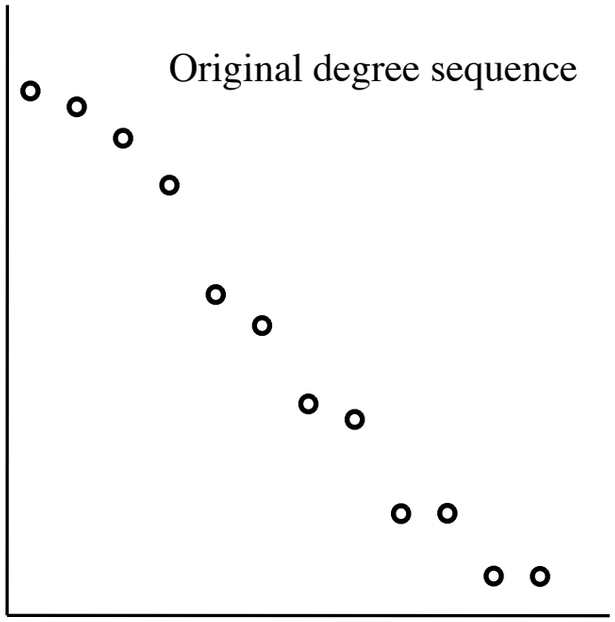
[100,100, 100, 98, 98,15,15,15]

[degree-sequence anonymization] Given degree sequence ***d***, and integer ***k***, construct ***k***-anonymous sequence ***d'*** such that ***||d'-d||*** **is minimized**

Increase/decrease of degrees correspond to additions/deletions of edges

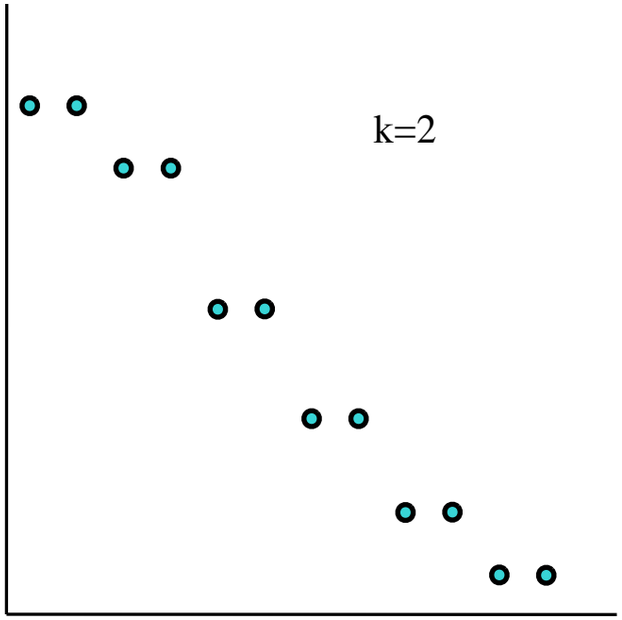
evimaria@cs.bu.edu

Algorithm for degree-sequence anonymization



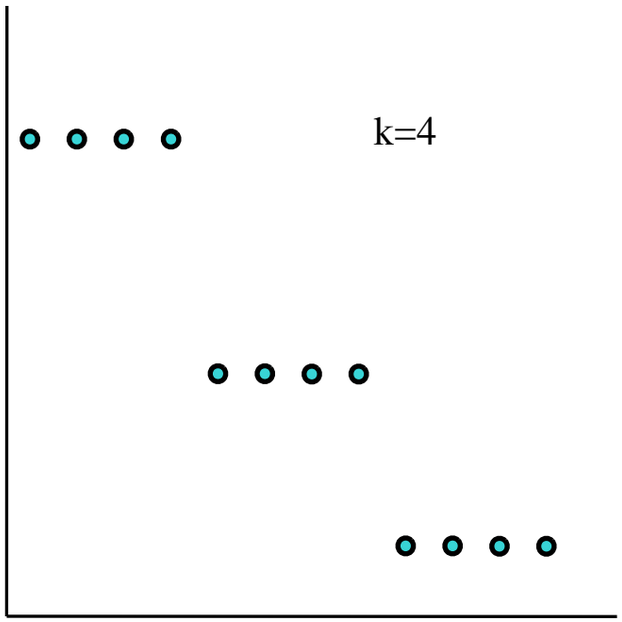
evimaria@cs.bu.edu

Algorithm for degree-sequence anonymization



evimaria@cs.bu.edu

Algorithm for degree-sequence anonymization



evimaria@cs.bu.edu

DP for degree-sequence anonymization

- ***d*** *(1)* ≥ ***d*** *(2)* ≥… ≥ ***d*** *(i)* ≥… ≥ ***d*** *(n)* : original degree sequence.
- ***d'*** *(1)* ≥ ***d'*** *(2)* ≥…≥ ***d'*** *(i)* ≥…≥ ***d'*** *(n)* : k-anonymized degree sequence.
- ***C(i, j)*** : anonymization cost when all nodes *i, i+1, …, j* are put in the same anonymized group, i.e.,

$$C(i,j) = \sum_{l=i}^{j} \left(d(i) - d^{*} \right)$$

- ***DA(1, n)*** : the optimal degree-sequence anonymization cost
- Dynamic Programming with **O(n^2)**

$$DA(1,i) = \min_{k \le t \le i-k} \{ DA(1,t) + C(t+1,i) \}$$

- Dynamic Programming with **O(nk)**

$$DA(1,i) = \min_{\max\{k, i-2k+1\} \le t \le i-k} \{ DA(1,t) + C(t+1,i) \}$$

- Dynamic Programming can be done in **O(n)** with some additional bookkeeping

evimaria@cs.bu.edu

GraphAnonymization algorithm

Input: Graph ***G*** with degree sequence ***d***, integer ***k***
Output: ***k***-degree anonymous graph ***G'***

[**Degree Sequence Anonymization**]:

- Contruct an anonymized degree sequence ***d'*** from the original degree sequence ***d***

[**Graph Construction**]:

[**Construct**]: Given degree sequence ***d'***, construct a new graph $G^0(V, E^0)$ such that the degree sequence of G^0 is ***d'***

[**Transform**]: Transform $G^0(V, E^0)$ to $G'(V, E')$ so that ***SymDiff***(G', G) is minimized.

evimaria@cs.bu.edu

Are all degree sequences realizable?

- A degree sequence ***d*** is **realizable** if there exists a simple undirected graph with nodes having degree sequence ***d.***

- Not all vectors of integers are realizable degree sequences
  - d = {4,2,2,2,1} ?

- How can we decide?

evimaria@cs.bu.edu

Realizability of degree sequences

[**Erdös and Gallai**] A degree sequence $\boldsymbol{d}$ with $\boldsymbol{d}(1) \geq \boldsymbol{d}(2) \geq \ldots \geq \boldsymbol{d}(i) \geq \ldots \geq \boldsymbol{d}(n)$ and $\Sigma \boldsymbol{d}(i)$ even, is realizable if and only if

$$\sum_{i=1}^{l} \boldsymbol{d}(i) \leq l(l-1) + \sum_{i=l+1}^{n} \min\{l, \boldsymbol{d}(i)\}, \text{ for every } 1 \leq l \leq n-1.$$

BOSTON UNIVERSITY

evimaria@cs.bu.edu

Realizability of degree sequences

[**Erdös and Gallai**] A degree sequence ***d*** with ***d***(*1*) ≥ ***d***(*2*) ≥… ≥ ***d***(*i*) ≥… ≥ ***d***(*n*) and Σ***d****(i)* even, is realizable if and only if

$$\sum_{i=1}^{l} \boldsymbol{d}(i) \le l(l-1) + \sum_{i=l+1}^{n} \min\{l, \boldsymbol{d}(i)\}, \text{ for every } 1 \le l \le n-1.$$

Input: Degree sequence ***d'***
Output: Graph ***$G^0(V, E^0)$*** with degree sequence ***d'*** or ***NO!***

→If the degree sequence ***d'*** is NOT realizable?

- Convert it into a realizable and ***k***-anonymous degree sequence

evimaria@cs.bu.edu

GraphAnonymization algorithm

Input: Graph ***G*** with degree sequence ***d***, integer ***k***
Output: ***k***-degree anonymous graph ***G'***

[**Degree Sequence Anonymization**]:

- Contruct an anonymized degree sequence ***d'*** from the original degree sequence ***d***

[**Graph Construction**]:

[**Construct**]: Given degree sequence ***d'***, construct a new graph $G^0(V, E^0)$ such that the degree sequence of G^0 is ***d'***

[**Transform**]: Transform $G^0(V, E^0)$ to $G'(V, E')$ so that ***SymDiff***(G', G) is minimized.

evimaria@cs.bu.edu

Graph-transformation algorithm

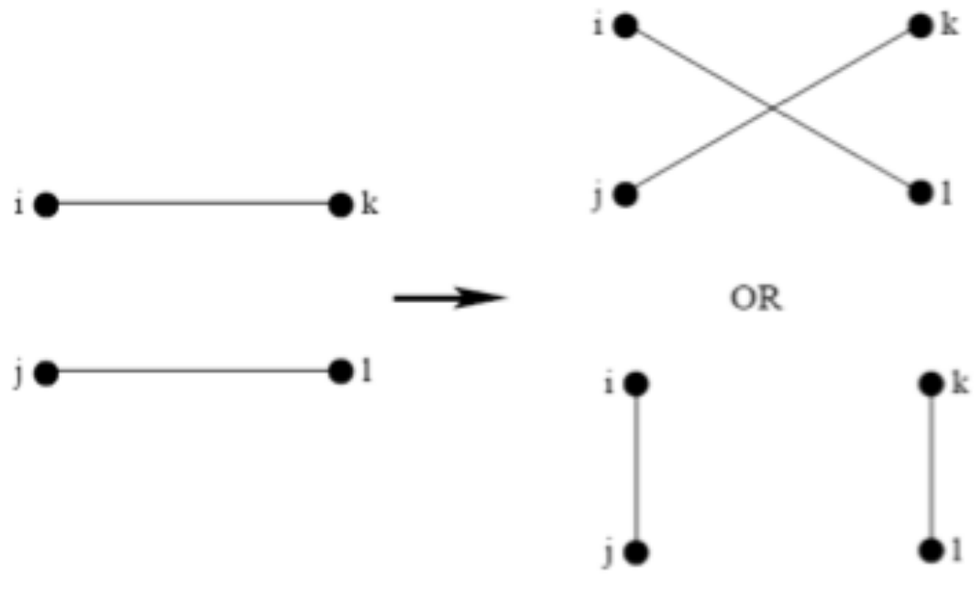
- GreedySwap transforms $G^0 = (V, E^0)$ into $G'(V, E')$ with the same degree sequence d', and min symmetric difference $SymDiff(G',G)$.

- GreedySwap is a greedy heuristic with several iterations.

- At each step, GreedySwap swaps a pair of edges to make the graph more similar to the original graph G, while leaving the nodes' degrees intact.

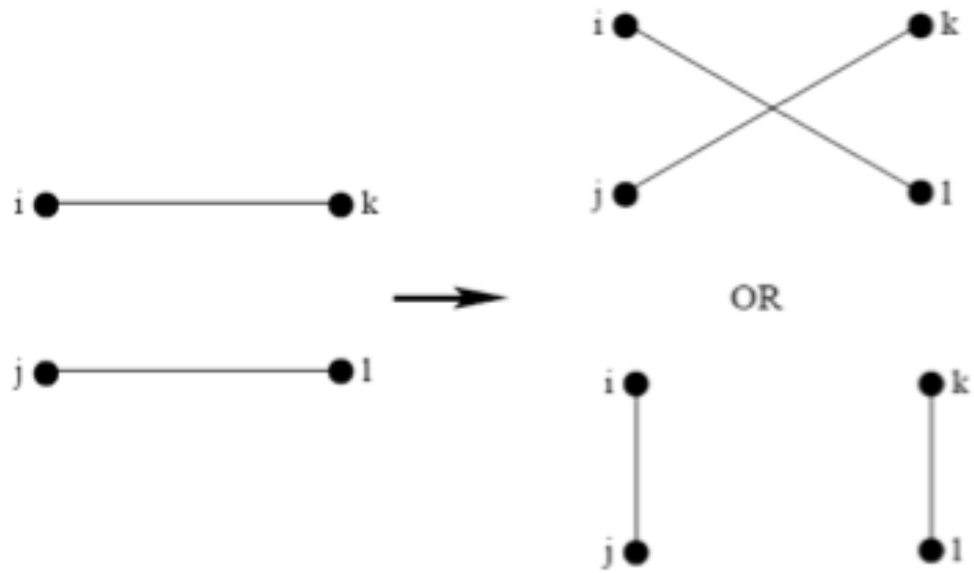
evimaria@cs.bu.edu

Valid swappable pairs of edges



evimaria@cs.bu.edu

Valid swappable pairs of edges



A swap is ***valid*** if the resulting graph is simple

evimaria@cs.bu.edu

GreedySwap algorithm

Input: A pliable graph $G^0(V, E^0)$, fixed graph $G(V,E)$
Output: Graph $G'(V, E')$ with the same degree sequence as $G^0(V,E^0)$

i=0
Repeat
find the valid swap in G^i that most reduces its symmetric difference with G , and form graph G^{i+1}
i++

BOSTON UNIVERSITY
evimaria@cs.bu.edu

Outline

- **Problem definition**
- **Algorithms**
- **Experimental results**

evimaria@cs.bu.edu

Experiments

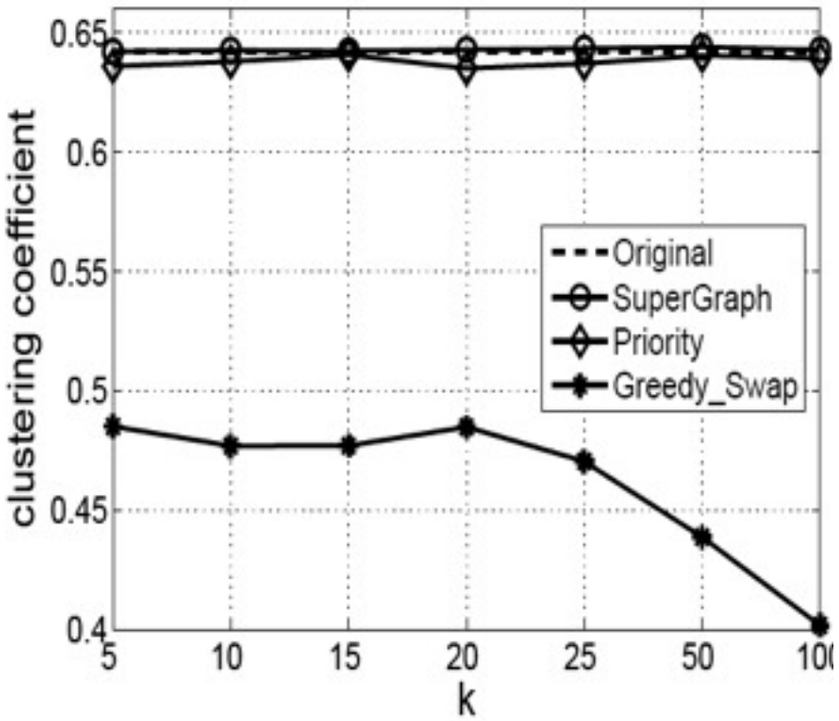
- **Datasets:** Co-authors, Enron emails, powergrid, Erdos-Renyi, small-world and power-law graphs

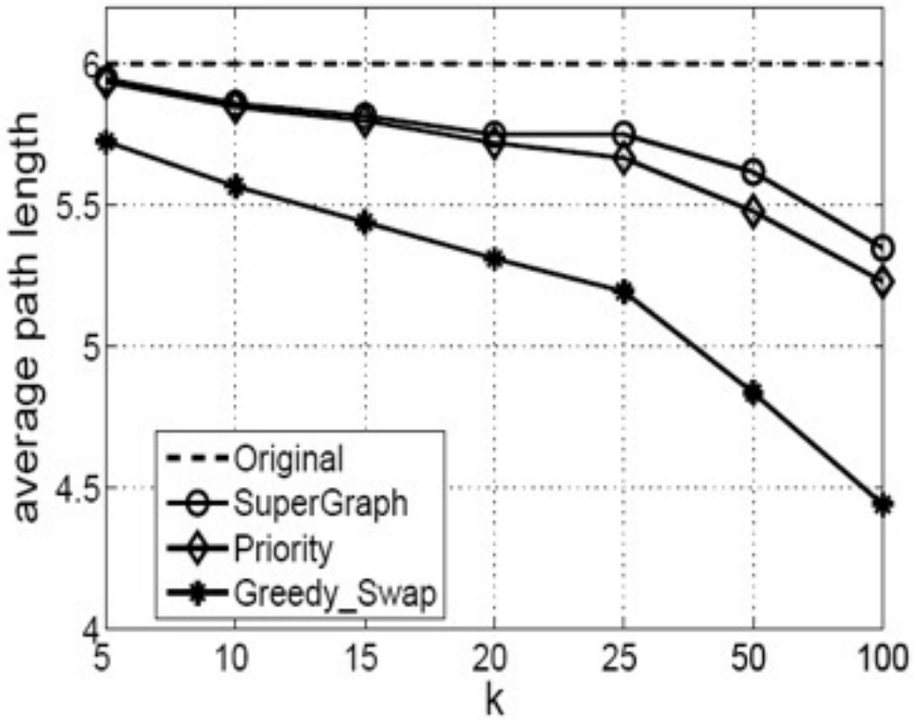
- **Goal:** degree-anonymization does not destroy the structure of the graph
  - Average path length
  - Clustering coefficient
  - Exponent of power-law distribution

evimaria@cs.bu.edu

Experiments: Clustering coefficient and Avg Path Length

- **Co-author** dataset
- APL and CC do not change dramatically even for large values of ***k***





evimaria@cs.bu.edu

Experiments: Edge intersections

Edge intersection achieved by the GreedySwap algorithm for different datasets.

Parenthesis value indicates the original value of edge intersection

Synthetic datasets	
Small world graphs*	0.99 (0.01)
Random graphs	0.99 (0.01)
Power law graphs**	0.93 (0.04)
Real datasets	
Enron	0.95 (0.16)
Powergrid	0.97 (0.01)
Co-authors	0.91(0.01)

(*) L. Barabasi and R. Albert: Emergence of scaling in random networks. *Science 1999.*
(**) Watts, D. J. Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology* 1999

evimaria@cs.bu.edu

Experiments: Exponent of power law distributions

Original	2.07
k=10	2.45
k=15	2.33
k=20	2.28
k=25	2.25
k=50	2.05
k=100	1.92

Co-author dataset

Exponent of the power-law distribution as a function of ***k***

evimaria@cs.bu.edu

Conclusions

- Problem and algorithmic aspects of degree-anonymization on graphs.

- Degree-anonymity does not destroy the graph structure in practice

BOSTON UNIVERSITY

evimaria@cs.bu.edu

Inverse anonymization problems

- Given two social networks that share a large portion of their nodes, can you map the nodes of the one network to the other?
- Examples: Twitter and FB. LinkedIn and FB etc.

evimaria@cs.bu.edu

Questions?

BOSTON UNIVERSITY

evimaria@cs.bu.edu

k-anonymity on tabular data

[k-Anonymity*] A dataset is ***k**-anonymous* if every record is indistinguishable from at least ***(k-1)*** other records.

[Algorithms] Replace specific values with more general, but

* P. Samarati and L. Sweeney, “Generalizing data to provide anonymity when disclosing information,” PODS 1998.

evimaria@cs.bu.edu

k-anonymity on tabular data

[k-Anonymity*] A dataset is ***k**-anonymous* if every record is indistinguishable from at least **(*k-1*)** other records.

[Algorithms] Replace specific values with more general, but

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
t_1	1	0	0	0	1	1	1	0
t_2	1	1	0	0	1	0	1	0
t_3	1	0	1	0	1	1	1	0
t_4	0	1	1	1	0	0	0	1
t_5	1	1	1	1	1	0	0	1
t_6	1	1	1	1	1	0	0	0

* P. Samarati and L. Sweeney, “Generalizing data to provide anonymity when disclosing information,” PODS 1998.

evimaria@cs.bu.edu

k-anonymity on tabular data

[k-Anonymity*] A dataset is ***k**-anonymous* if every record is indistinguishable from at least **(*k-1*)** other records.

[Algorithms] Replace specific values with more general, but

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
t_1	1	*	*	0	1	*	1	0
t_2	1	*	*	0	1	*	1	0
t_3	1	*	*	0	1	*	1	0
t_4	*	1	1	1	*	0	0	*
t_5	*	1	1	1	*	0	0	*
t_6	*	1	1	1	*	0	0	*

* P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," PODS 1998.

evimaria@cs.bu.edu

Anonymizing adjacency matrices

BOSTON UNIVERSITY

evimaria@cs.bu.edu

Anonymizing adjacency matrices

All graphs are unweighted and undirected

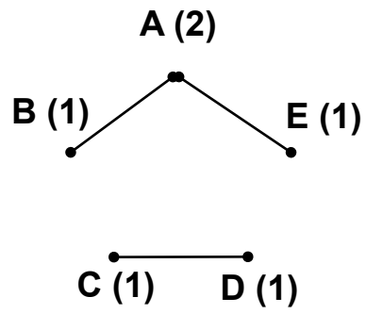
BOSTON UNIVERSITY

evimaria@cs.bu.edu

Anonymizing adjacency matrices

Original Graph

	A	B	C	D	E
A	0	1	0	0	1
B	1	0	0	0	0
C	0	0	0	1	0
D	0	0	1	0	0
E	1	0	0	0	0



All graphs are unweighted and undirected

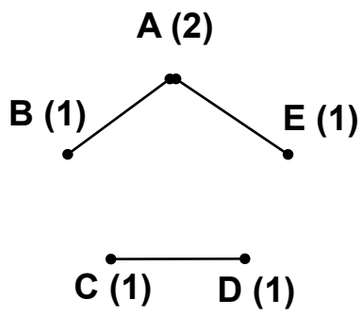
BOSTON UNIVERSITY

evimaria@cs.bu.edu

Anonymizing adjacency matrices

Original Graph

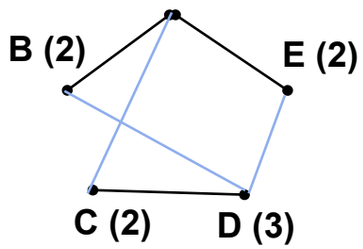
	A	B	C	D	E
A	0	1	0	0	1
B	1	0	0	0	0
C	0	0	0	1	0
D	0	0	1	0	0
E	1	0	0	0	0



All graphs are unweighted and undirected

2-anonymous graph based on k-anonymity model.

	A	B	C	D	E
A	0	1	1	0	1
B	1	0	0	1	0
C	1	0	0	1	0
D	0	1	1	0	1
E	1	0	0	1	0



1. Changes to one entry leads to changes to another.

2. k-anonymity could result in significant changes of the graph and would make the graph useless.

evimaria@cs.bu.edu

Experiments: Degree-sequence anonymizati

- Co-author dataset

- Degree sequences do not change dramatically even for large values of ***k***

evimaria@cs.bu.edu