IEEE INFOCOM 2013, April 14-19, Turin, Italy

Outsourcing Privacy-Preserving Social Networks to a Cloud

Guojun Wang[a], Qin Liu[a], **Feng Li[c]**, Shuhui Yang[d], and Jie Wu
b

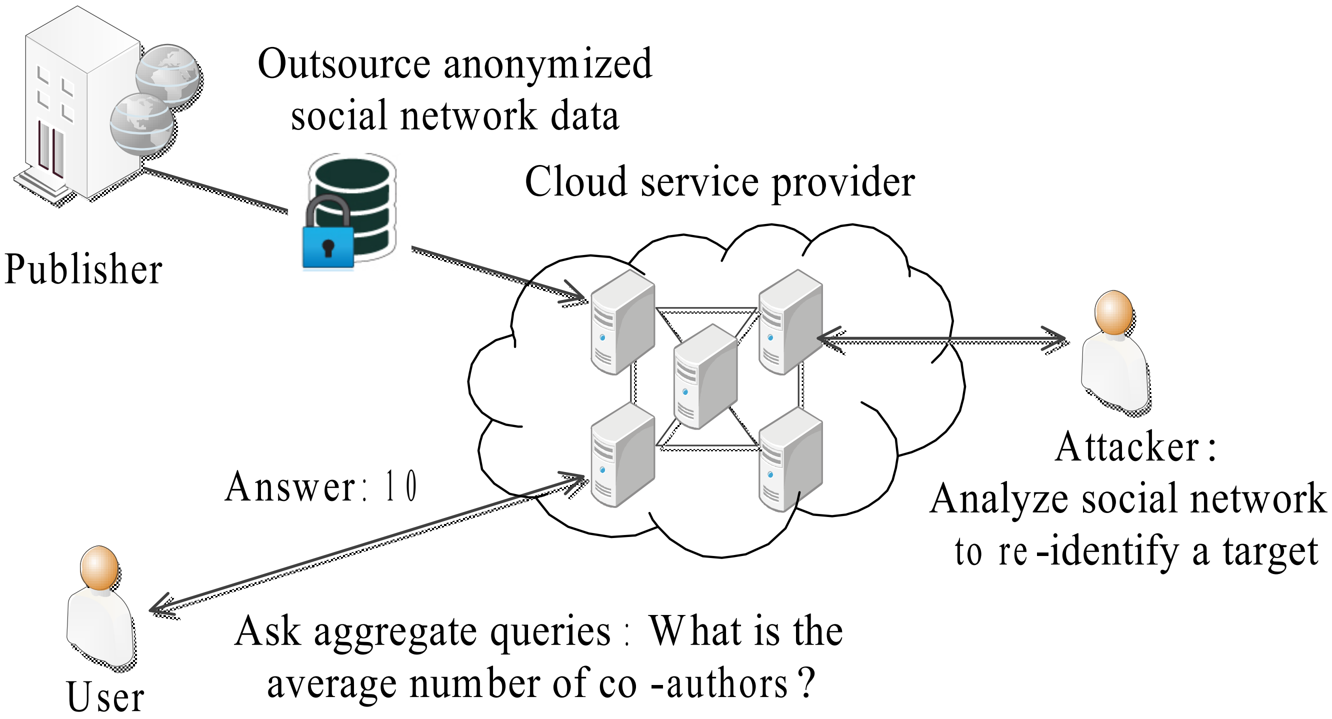
[a] Central South University, China
[b] Temple University, USA
[c] Indiana University-Purdue University Indianapolis, USA
[d] Purdue University Calumet, USA

2013-4-18

Introduction

Cloud Computing Model

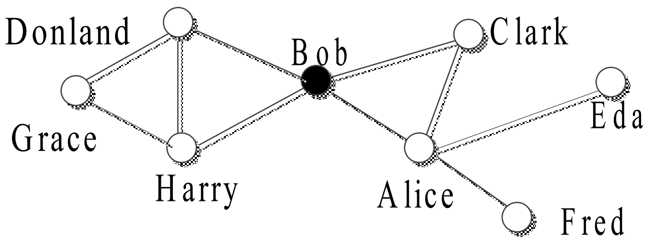
- Cloud computing as a new commercial paradigm enables organizations that host social network data to outsource a portion of their data to a cloud.



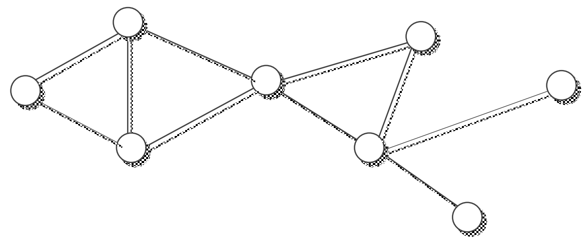
How to protect individuals' identities?

1-Neighborhood Attack

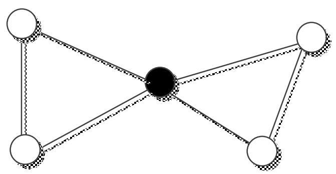
- Anonymization cannot resist the 1-neighborhood attack, where the attacker is assumed to know the target's 1-neighborhood graph.



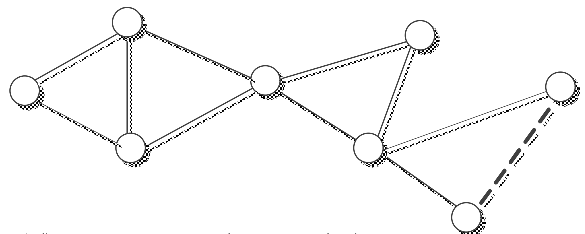
(a) Initial social network



c) Anonymized social network



(b) 1-neighborhood of Bob

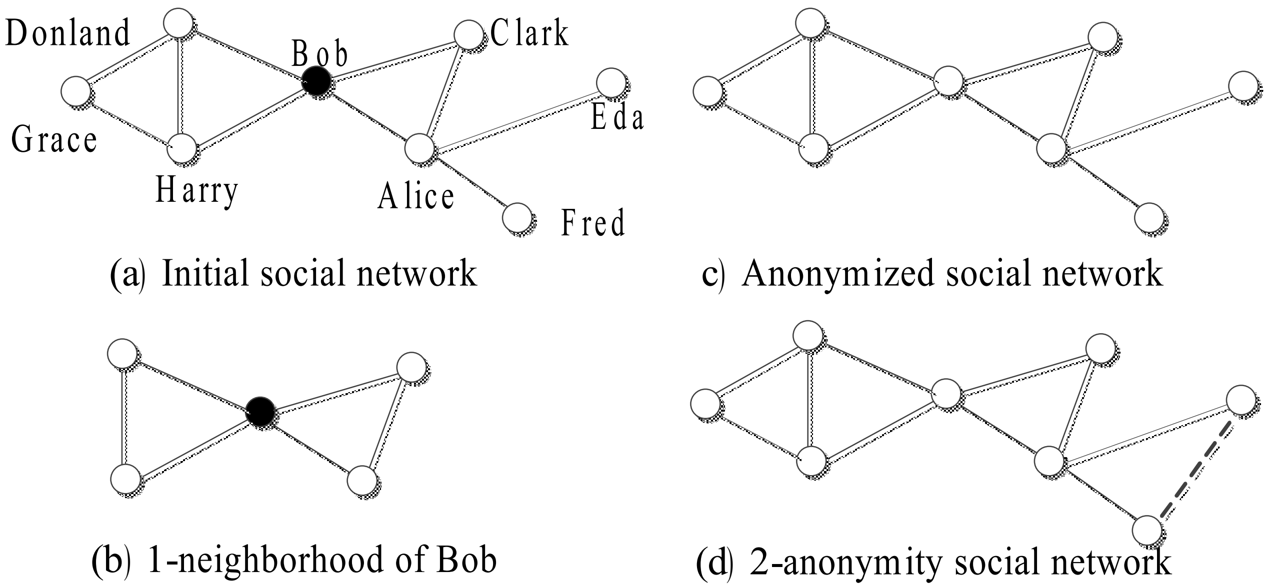


(d) 2-anonymity social network

Existing work made any node's 1-neighborhood graph isomorphic with at least k − 1 other nodes' graphs by adding noise edges (k-anonymity).

Challenges

- K-anonymity cannot resist 1*-neighborhood attack, where an attacker is assumed to know the degrees of the target's one-hop neighbors, in addition to the 1-neighborhood graph.



(a) Initial social network

c) Anonymized social network

(b) 1-neighborhood of Bob

(d) 2-anonymity social network

Existing work requires the addition of more edges, so that the degrees of the K-isomorphic graphs are the same. The utility of the graph is reduced.

Our Contributions

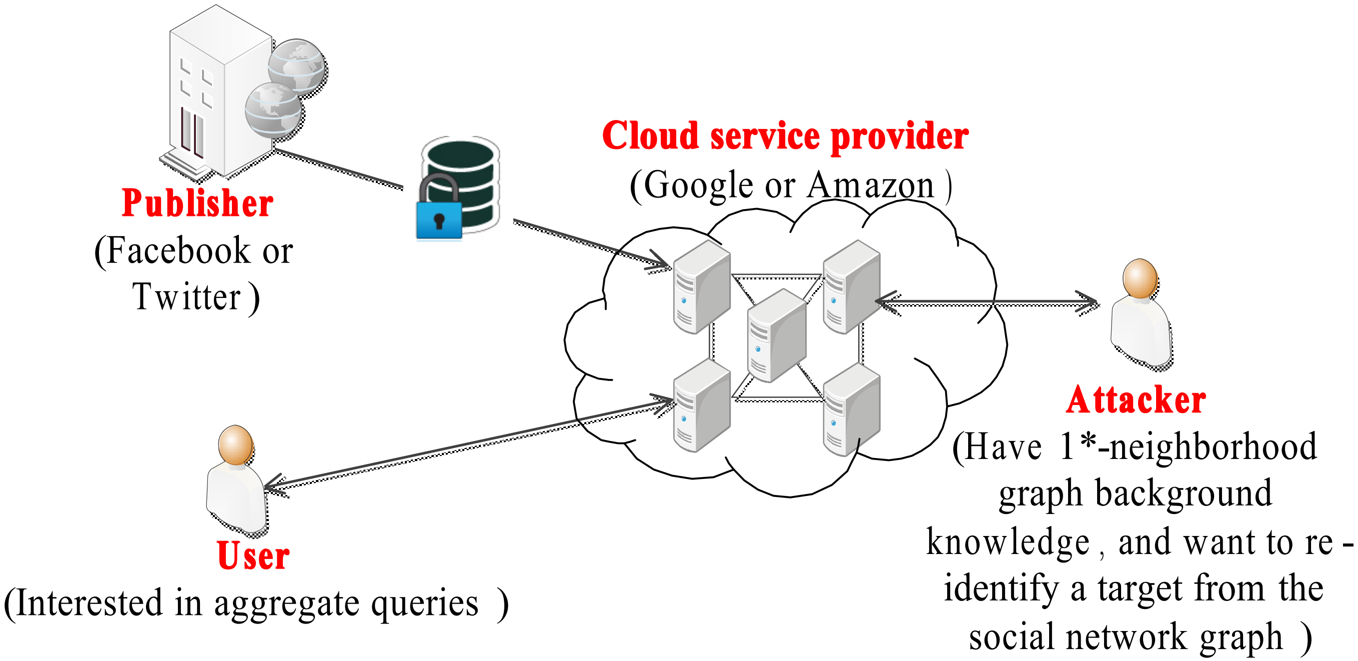
- We identify a novel attack, 1*-neighborhood attack, for outsourcing social networks to a cloud.

- We define the probabilistic indistinguishability property for an outsourced social network, and propose a heuristic indistinguishable group anonymization scheme (HIGA) to generate social networks with this privacy property.

- We conduct experiments on both synthetic and real data sets to verify the effectiveness of the proposed scheme.

Preliminaries

System Model



Privacy goal. Given any target's 1-*neighborhood graph, the attacker cannot re-identify the target from an anonymized social network with confidence higher than a threshold.

Utility goal. The anonymized social networks can be used to answer aggregate queries with high accuracy.

Problem Formulation

Problem Definition. *Given a network graph* $\mathcal{G} = (V(\mathcal{G}), E(\mathcal{G}))$ *and a positive integer* k*, derive an anonymized graph* $\mathcal{G}' = (V(\mathcal{G}'), E(\mathcal{G}'))$ *to be published, such that (1)* $V(\mathcal{G}') = V(\mathcal{G})$*; (2)* $\mathcal{G}'$ *is probabilistically indistinguishable with respect to* $\mathcal{G}$*; (3) the anonymization from* $\mathcal{G}$ *to* $\mathcal{G}'$ *has minimal anonymization cost.*

The problem of generating a social network with above three properties is NP-hard.

Definitions

- Let $G^*{}_u$ and $G'^*{}_u$ denote the 1*-neighborhood graph of node u in the original social network G and in the anonymized social network G', respectively.

Node Indistinguishability. *Nodes u and v are indistinguishable if an observer cannot decide whether or not $G_u^* \neq G_v^*$ in the original graph $\mathcal{G}$, by comparing $G_u'^*$ and $G_v'^*$ in an anonymized graph $\mathcal{G}'$.*

Group Indistinguishability. *For a group of nodes $g = \{v|v \in V(\mathcal{G})\}$ and $|g| \geq k$ if for each pair of nodes $\{\langle u, v\rangle | u, v \in g\}$, u and v are indistinguishable in the published graph $\mathcal{G}'$, group g is an indistinguishable group.*

Probabilistic Indistinguishability. *A published social network $\mathcal{G}'$ achieves probabilistic indistinguishability, if all nodes $\{v|v \in V(\mathcal{G}')\}$ can be classified into $m \geq 1$ groups, where each group has the property of* group indistinguishability.

Scheme Description

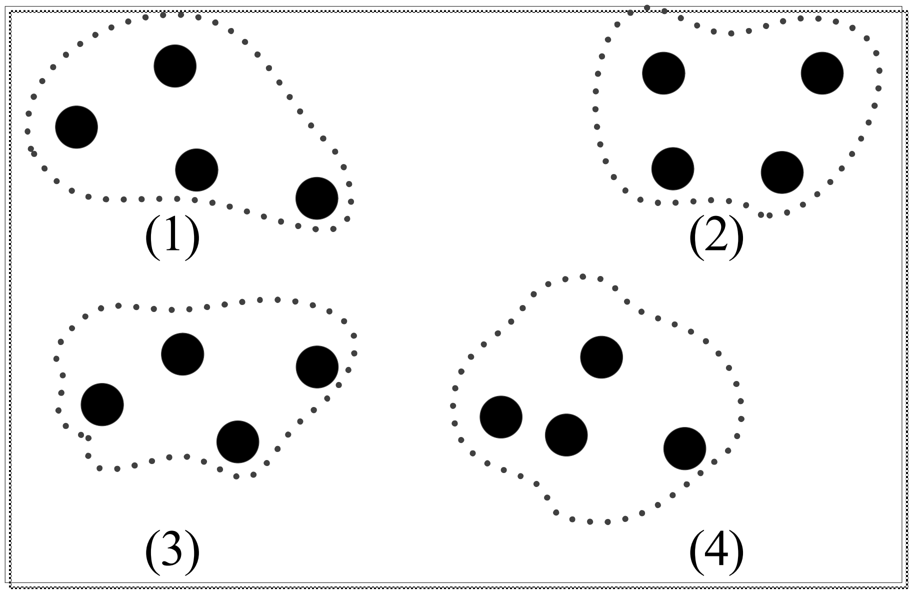
Intuition

The heuristic indistinguishable group anonymization (HIGA) scheme consists of 4 steps:

- Grouping
- Testing
- Anonymization
- Randomization

Intuition

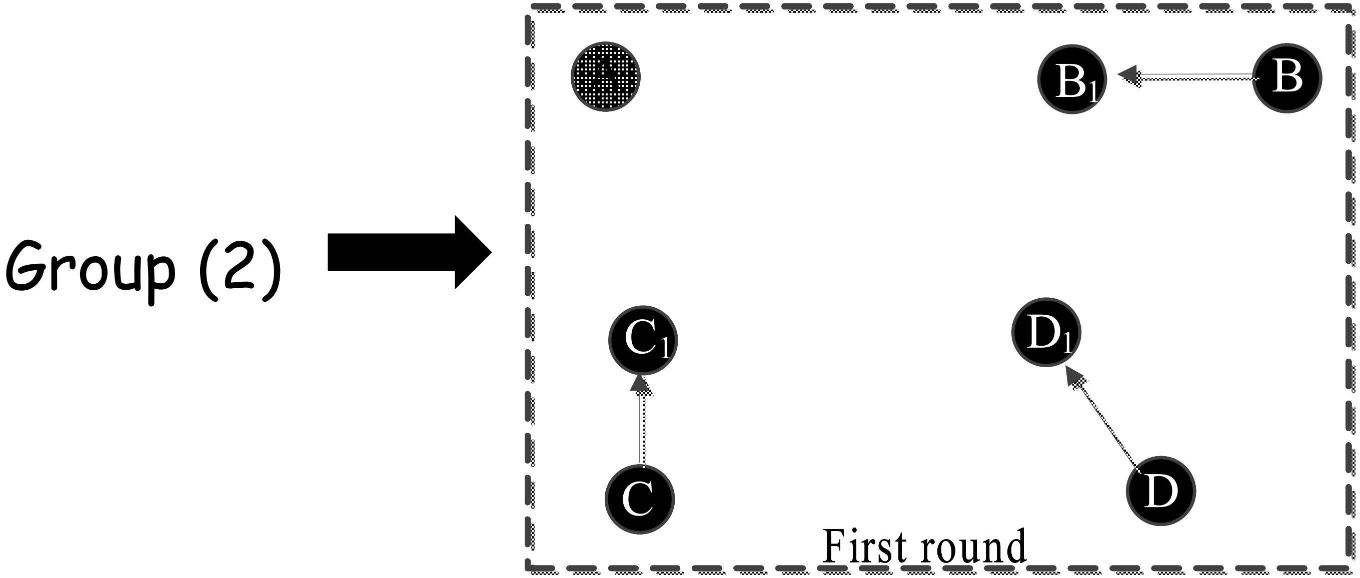
Grouping classifies nodes whose 1*-neighborhood graphs satisfy certain metrics into groups, where each group size is at least equal to k.



(A) Grouping

Intuition

Testing uses random walk (RW) to test whether the 1-neighborhood graphs of nodes in a group approximately match or not.



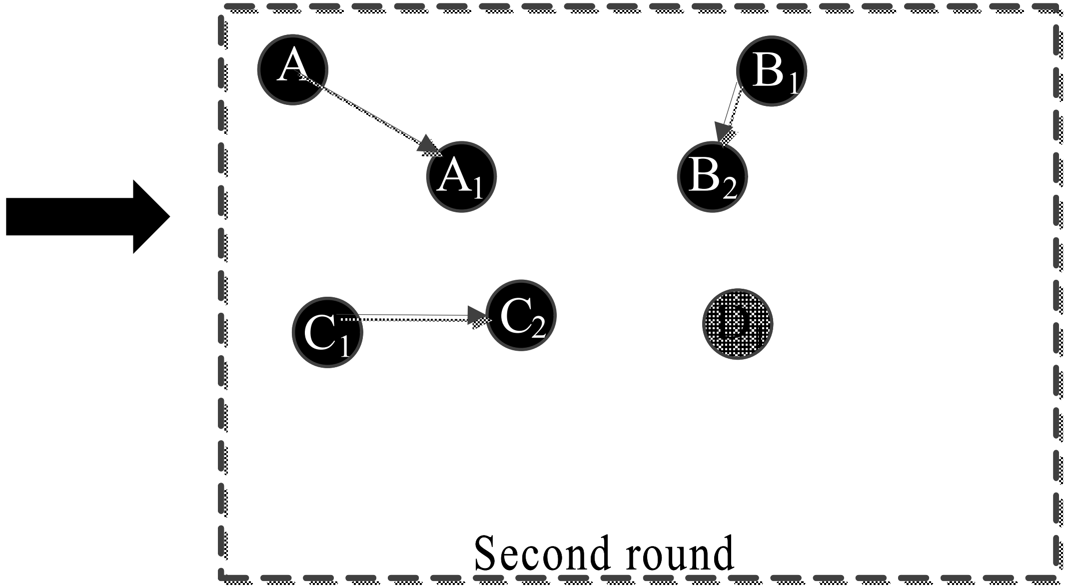
(B) Testing and anonymization

Anonymization uses a heuristic anonymization algorithm to make the 1-neighborhood graphs of nodes in each group approximately match

Intuition

Testing uses random walk (RW) to test whether the 1-neighborhood graphs of nodes in a group approximately match or not.

Group (2)

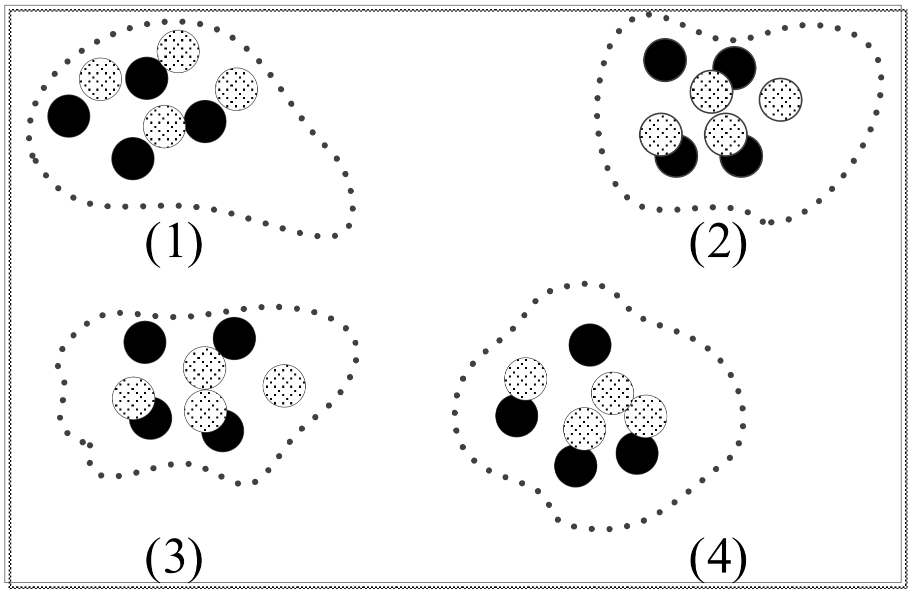


(B) Testing and anonymization

Anonymization uses a heuristic anonymization algorithm to make the 1-neighborhood graphs of nodes in each group approximately match

Intuition

Randomization randomly modifies the graph with certain probability to make each node's 1*-neighborhood graph be changed with certain probability



(C) Randomization

Step 1: Grouping

A social network is modeled as an undirected and unlabeled graph $G = (V(G), E(G))$, where $V(G)$ is a set of nodes, and $E(G)$ $\quad V(G) \times V(G)$ is a set of edges.

1-Neighborhood Graph. $G_u = (V_u, E_u)$, *where* V_u *denotes a set of nodes* $\{v|(u, v) \in E(\mathcal{G}) \vee (v = u)\}$, *and* E_u *denotes a set of edges* $\{(w, v)|(w, v) \in E(\mathcal{G}) \wedge \{w, v\} \in V_u\}$.

1*-Neighborhood Graph. $G_u^* = (G_u, D_u)$, *where* G_u *is the 1-neighborhood graph of node u, and* D_u *is a sequence of degrees of u's one-hop neighbors.*

Step 1: Grouping

We group nodes by using the following metric: number of one-hop neighbors, in-degree sequence, out-degree sequence, total number of edges, and betweenness.

In-degree sequence. $I_v = \{|E_u^+|\}_{u \in V_v}$, *where* $E_u^+ = \{(u,w)|w \in V_v\}$, *and* $|E_u^+|$ *is the number of edges in* E_u^+.

Out-degree sequence. $O_v = \{|E_u^-|\}_{u \in V_v}$, *where* $E_u^- = \{(u,w)|w \notin V_v\}$, *and* $|E_u^-|$ *is the number of edges in* E_u^-.

Betweenness. $B_v = |V_v^*|/|V_v^+|$, *where* $V^* = \{\langle u,w\rangle|u,v \in V_v \wedge (u,w) \notin E_v\}$, *and* $V_v^+ = \{\langle u,w\rangle|u,v \in V_v\}$.

TABLE I

EFFECTIVENESS OF METRICS FOR GROUPING

Nodes	Edges	Percentage
$100 \sim 200$	$1,000 \sim 2,000$	47%
$500 \sim 1,000$	$5,000 \sim 10,000$	61%

Step 2: Testing

We analyze each pair of nodes u and v by computing the steady states of their 1-neighborhood graphs G_u and G_v with RW.

$$p_{u_j}(t) = \sum_{u_i \in V(\mathcal{G})} \frac{1}{|V(\mathcal{G})|} \cdot (1-d) \cdot p_{u_i}(t-1) + \sum_{u_i \in N(u_j)} \frac{1}{|N(u_i)|} \cdot d \cdot p_{u_i}(t-1) \quad (1)$$

Eq. 1 calculates the probability of a node u_j being located at time t

$$\mathbf{p}(t) = \frac{(1-d)}{N} \cdot \mathbb{I} + d \cdot \mathbf{W} \cdot \mathbf{p}(t-1) \quad (2)$$

Eq. 2 calculates the probability distribution on all nodes in the graph

$$\mathbf{p}^{\star} = \frac{(1-d)}{|V(\mathcal{G})|} \cdot \sum_{k=0}^{\infty} d^k W^k \cdot \mathbb{I} \quad (3)$$

Eq. 3 calculates the steady state of Eq. 2

Step 2: Testing

We use Eq. 4 to calculate the Euclidean distance between the topological signatures of the nodes:

$$cost(x, w) = \sqrt{(\mathrm{p}_x^\star - \mathrm{p}_w^\star)^2} \qquad (4)$$

The cost for matching two 1-neighborhood graphs is calculated with Eq. 5

$$cost(G_u, G_v) = \sqrt{\sum_{x,w \notin \mathbb{V}} (\mathrm{p}_x^\star - \mathrm{p}_w^\star)^2 + (|\mathbb{V}| * \beta)} \qquad (5)$$

Approximate matching. *Let $G_u = (V(G_u), E(G_u))$ and $G_v = (V(G_v), E(G_v))$ be two graphs. G_u and G_v approximately match, denoted as $G_u \approx G_v$, if an optimal bipartite graph matching exists between $V(G_u)$ and $V(G_v)$, such that the $cost(G_u, G_v)$ is smaller than a threshold value α.*

How to decide α is the key problem

Step 3: Anonymization

Algorithm 1 Heuristic Anonymization Algorithm

{Given m groups $g_1, \ldots, g_m$ as CGS}
Sort CGS in descending order of the number of neighbors
while CGS is not empty **do**
  Choose the first group in CGS as the processing group g_* and remove g_* from CGS
  for each node u in g_* **do**
    Construct 1-neighborhood graph G_u
    Use Eq. 3 to calculate G_u's topological signatures
  for each pair of nodes (u, v) in g_* **do**
    Use Eq. 5 to calculate cost of matching G_u and G_v
  while exists a cost larger than α **do**
    Randomly choose a node $u \in g_*$ as the group seed
    for each node $v \in g_*$ **do**
      if $cost(G_u, G_v) > \alpha$ **then**
        Approach G_u to G_v with probability q
        Approach G_v to G_u with probability $1 - q$

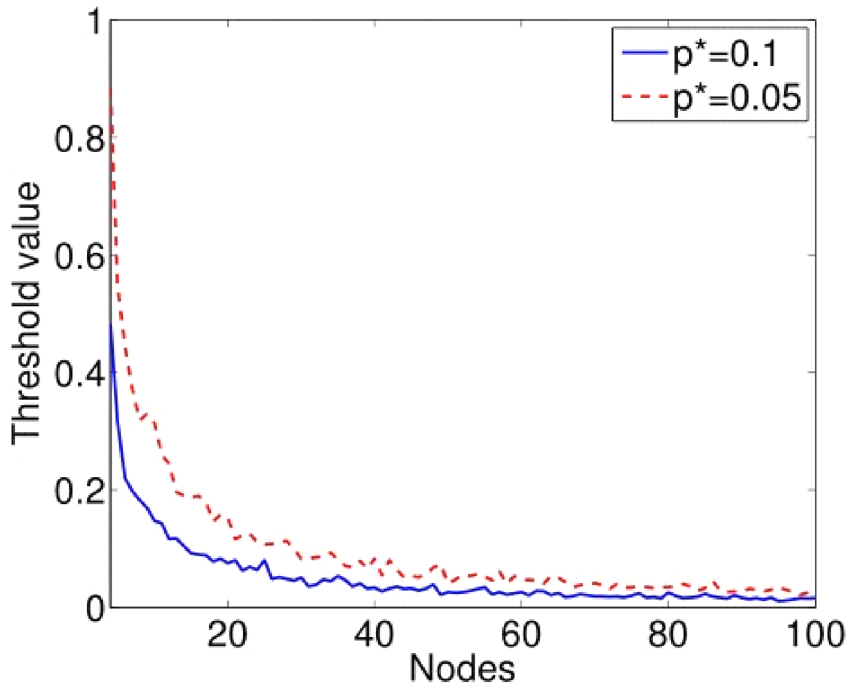
Step 4: Randomization

Given a randomization probability p. We first randomly remove p(|E(G)|) edges from the graph, and then for two nodes that are not linked, we add an edge with probability p.

The key problem lies in determining p to randomize the graph

Evaluation

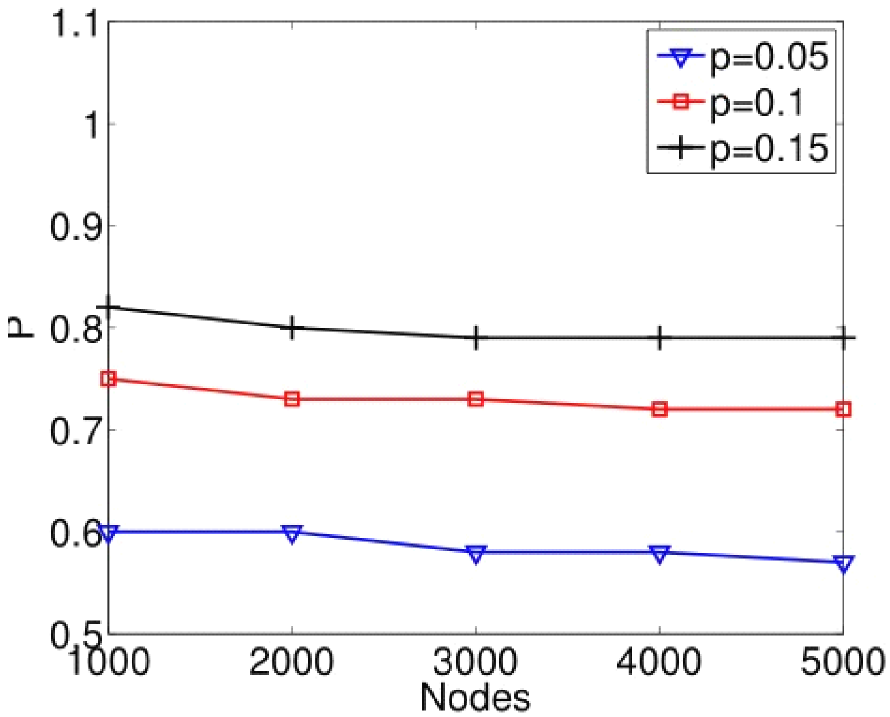
Parameter Setting



Threshold value α

- First, randomly generate a 1-neighborhood graph with N nodes
- Then generate a similar graph by randomly modifying p* percentage of edges

Parameter Setting

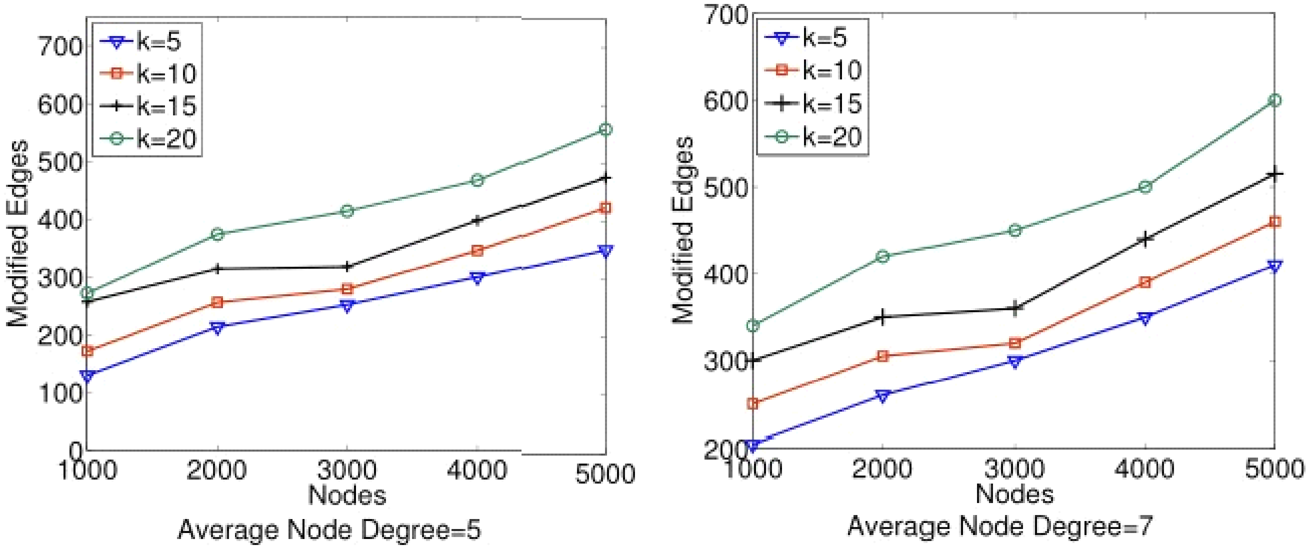


Random probability p

- First randomly generate a graph with N nodes and M edges.
- Then, randomize the graph with different p values, and calculate the percentage P of 1*-neighborhood graphs being changed in the randomized graph.

Synthetic Data Set

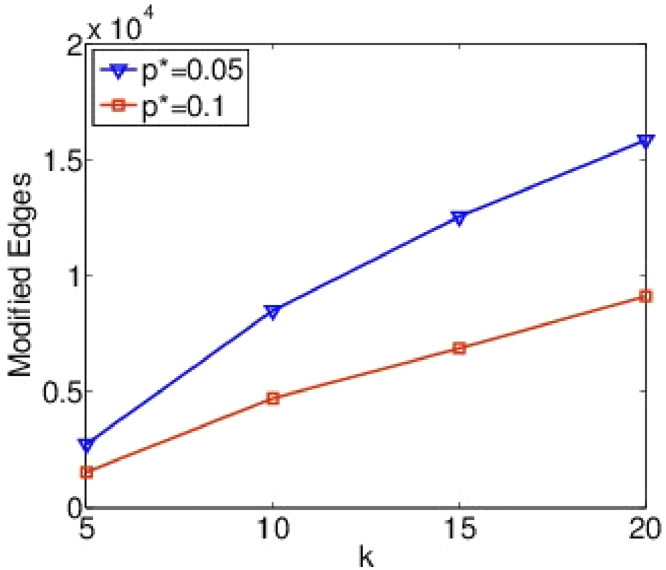
* We use the Barabai-Albert algorithm (B-A algorithm) to generate synthetic data sets.
* First generate a network of a small size (5 nodes), and then use that network as a seed to build a larger-sized network (1,000, 2,000, 3,000, 4,000, and 5,000 nodes).



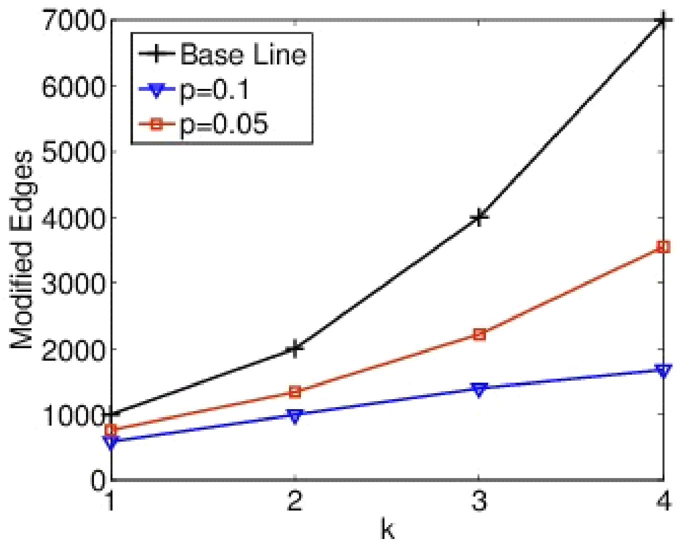
Number of modified edges on synthetic data sets. p∗ = 0.1.

Real Data Set

- Real social network, Astro Physics collaboration network, which contains 18,772 nodes and 396,160 edges. If an author i co-authored a paper with author j, the graph contains an undirected edge from i to j.



(a) Number of modified edges



(b) Comparison with existing work

Real Data Set

- The maximal node degree: MAX
- The minimal node degree: MIN
- The average node degree: AVE
- The error rate for answering the shortest distance queries: Error Rate

TABLE II
USABILITY OF THE ANONYMIZED SOCIAL NETWORK

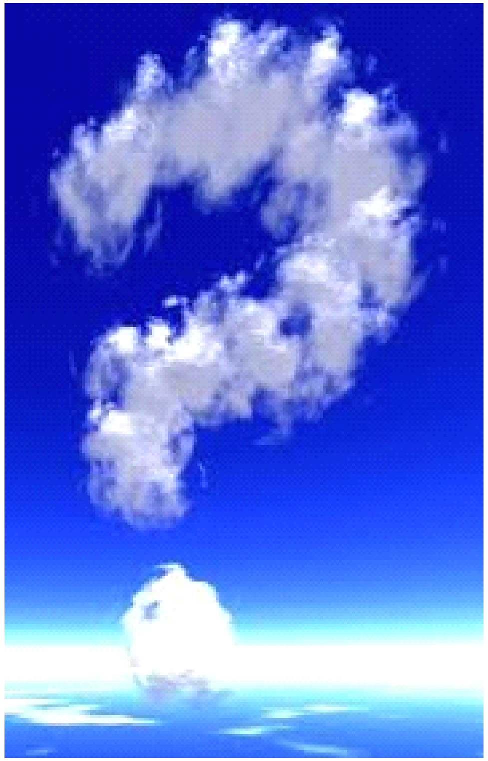
	Max	MIN	AVE	Error Rate
Original	505	2	22.1	0
k=5	505	2	22.4	2.9%
k=10	496	2	22.6	6.4%
k=15	485	2	22.9	8.1%
k=20	476	2	23.3	8.3%

Conclusion

We identify a novel 1*-neighborhood attack for publishing a social network graph to a cloud

We define a key property, probabilistic indistinguishability, for anonymizing outsourced social networks

We propose a heuristic anonymization scheme to anonymize social networks with this property



Thank you!