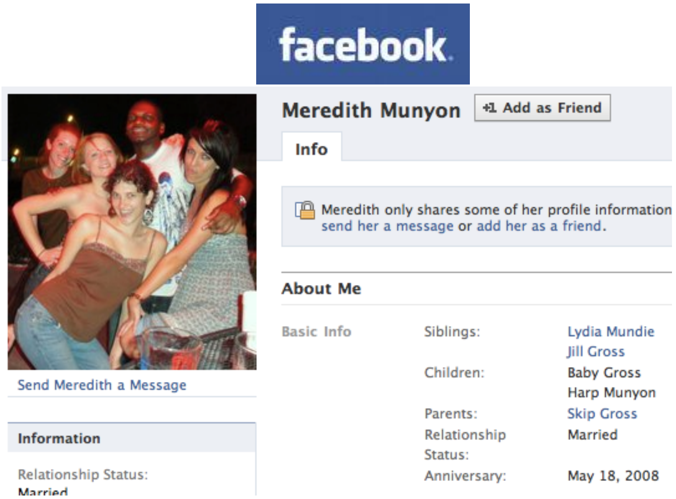
Prediction Promotes Privacy in Dynamic Social Networks

Smriti Bhagat (Rutgers), Graham Cormode,

Balachander Krishnamurthy and Divesh Srivastava
(AT&T Labs—Research)

Social Network Data

- Online Social Networks (OSNs) encode a variety of information about users and their interactions
- Not all information is publicly available; users choose its visibility
- Researchers want access to OSN data to
  - analyze communication patterns of a subpopulation
  - study users and their interactions over time
- Data owners want to share OSN data with third parties for
  - targeted advertisement
  - recommendations (yelp, pandora)



Meredith Munyon

Meredith only shares some of her profile information
send her a message or add her as a friend.

Send Meredith a Message

Information

Relationship Status:
Married

About Me

Basic Info	Siblings:	Lydia Mundie Jill Gross
	Children:	Baby Gross Harp Munyon
	Parents:	Skip Gross
	Relationship Status:	Married
	Anniversary:	May 18, 2008

Profile:
contact, personal info (friends, relationship, interests, affiliations)
Interaction:
Wall posts, private message, tag in photo

Anonymization of OSN Data

- Goal: publish a version of the social network data that preserves the privacy of users, yet is usable for analysis
- Simple example: Replace usernames by arbitrary IDs

Alex only shares some of his profile information with every
Alex, send him a message or add him as a friend.

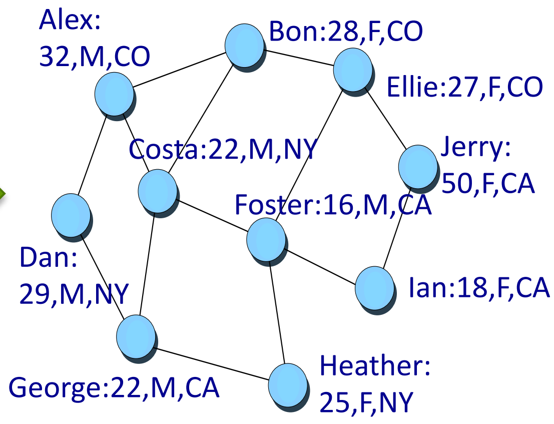
Networks:
South Island School '08

Relationship Status:
Single

Jerry only shares some of his profile information with every
a message or add him as a friend.

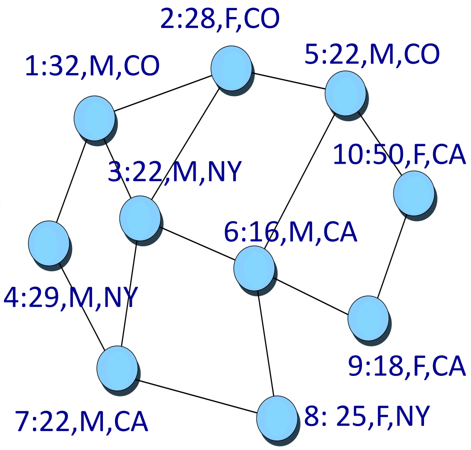
Basic Info Sex: Male
Siblings: Joey Price

Social Network Data



Graph representation of OSN:

- nodes are users with attributes (age, gender, location)
- edges represent a friendship



Anonymized graph: users are assigned arbitrary IDs

ID	User
1	Alex
2	Bon
…	

Anonymization: Recent work

- Two types of approaches:
  - Add/remove edges to make neighborhoods similar [Zhou et al. ICDE'08; Liu et al. SIGMOD'08]
  - Group users to hide user→node mapping [Hay et al. VLDB'08; Cormode et al. VLDB'08; Bhagat et al. VLDB'09]
- Focus has been on static network anonymization: publish a single anonymized instance of a social network
- Issue: insufficient for studying temporal trends in OSNs

This work

- Focus on dynamic network anonymization: repeatedly publish anonymized OSN data as the network evolves
- Challenge: Updates published should maintain the privacy of users
- Key contributions:
  - define privacy requirements for dynamic anonymization
  - propose an anonymization technique based on prediction
  - empirically evaluate privacy and utility over anonymized temporal data from Facebook, Flickr and Friendfeed

Dynamic Anonymization Problem

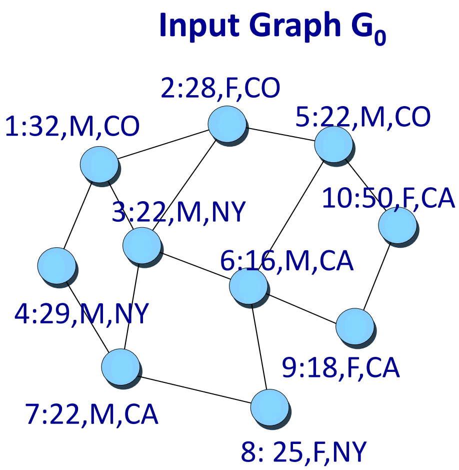
Input: $G_0, G_1 \ldots G_t$

$G_t=(V_t, U_t, E_t)$

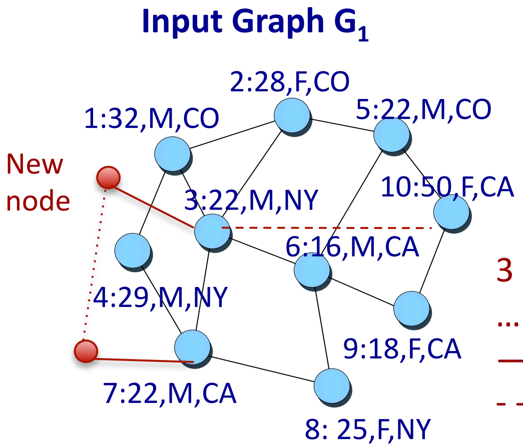
V_t nodes

U_t users

E_t edges



Input Graph G_0



Input Graph G_1

3 types of new edges (v,w)

… $v,w \in V_t \setminus V_{t-1}$

— $v \in V_{t-1}$, $w \in V_t \setminus V_{t-1}$

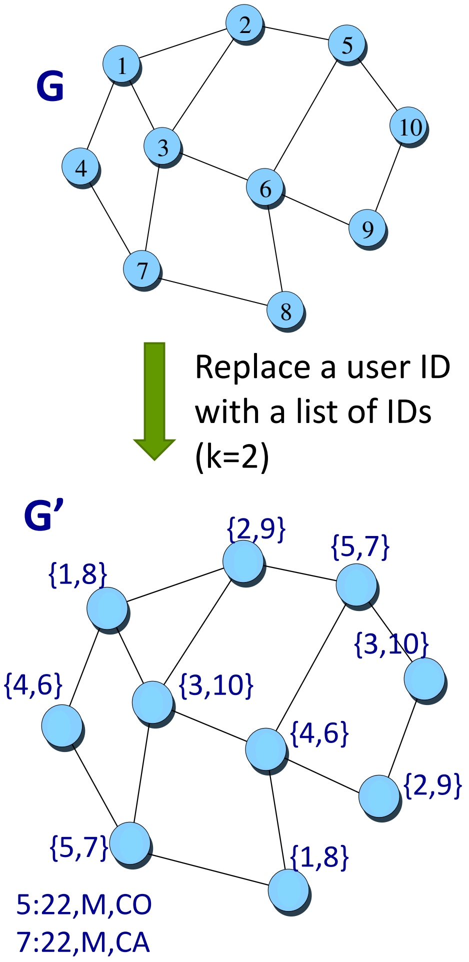
- - $v,w \in V_{t-1}$

Output: $G'_0, G'_1 \ldots G'_t$ with following properties:

- Privacy: Given G'_0, G'_1, …, G'_t
  - likelihood of identifying the node v in G'_t that represents user u is small
  - likelihood of determining if a pair of users (u_1,u_2) interact is small
- Utility: accurately answer queries
  - on the graph at any time t
  - over multiple instances of the network

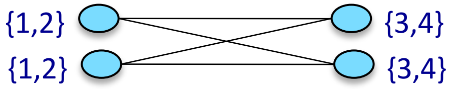
Note: For t=0, the problem definition reduces to that for static networks

Review: List-based Approach for Static Networks



Full List Approach

- Idea: Assign a list of userIDs to each node
- Approach:
  - Divide users into groups of size *k*, such that inter-group interactions are sparse
  - Grouping Condition: each user must interact with at most one user in any group



- Output: Publish full graph structure
- Properties:
  - Privacy: Given G’, an adversary can guess a user→node mapping, or an interaction between two users, with probability at most 1/k
  - Utility: Queries on graph properties answered exactly

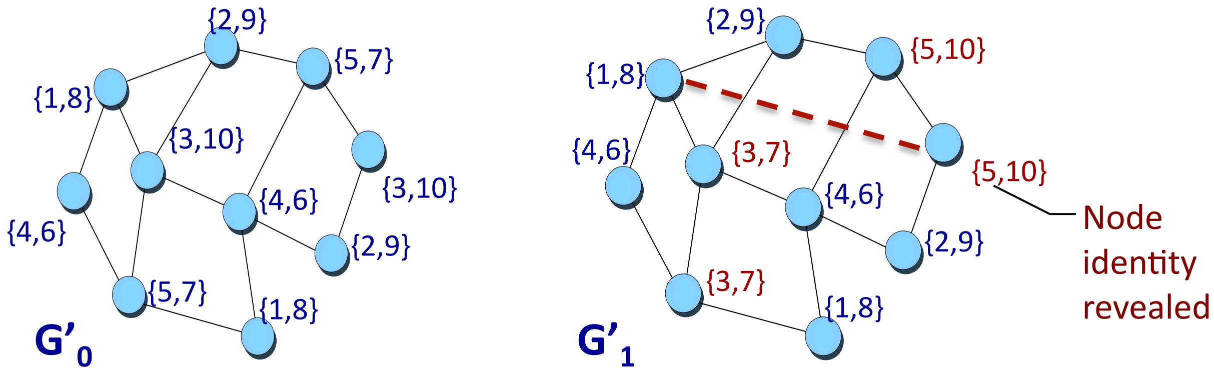
S. Bhagat, G. Cormode, B. Krishnamurthy, D. Srivastava. Class based graph anonymization for social network data, VLDB 09.

Dynamic Anonymization: Two Simple Approaches

Input: G_0 and G_1

Approach 1:

Individually anonymize instances using list-based approach



Independently anonymizing G_0 and G_1 leaks information

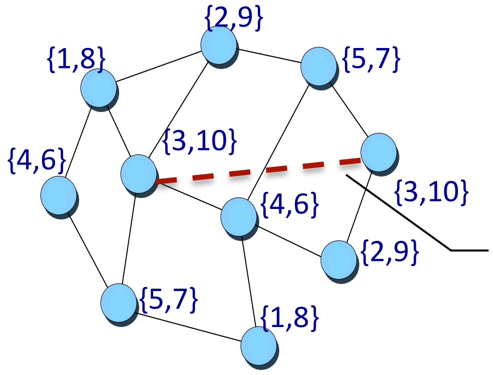
Must preserve initial anonymization

Dynamic Anonymization: Two Simple Approaches (cont.)

Input: G_0 and G_1

Approach 2: Naïve approach

Anonymize G'_0 and insert new edges as they appear



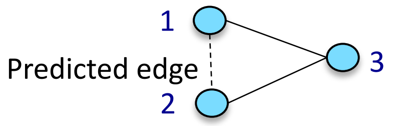
New edge reveals that users 3 and 10 interact

Addition of new edges may reduce sparsity in interactions among groups, thus weakening privacy

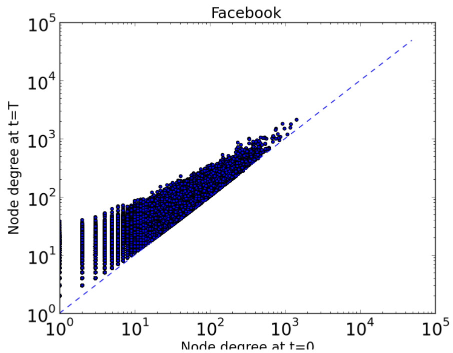
Group users such that interactions between groups are likely to be sparse on addition of new edges

Link Prediction

- Prediction Models:
  - Friend-of-a-friend (FOAF): unweighted
  - Common Neighbor (CN): w(i,j) = #common neighbors
  - Adamic-Adar (AA): w(i,j) = $\Sigma_c 1/\log(\deg(c))$, *c* is a CN
- Challenge:
  - Prediction models over-predict edges
    - Example: A regional sample of Facebook graph of 54K nodes, FOAF predicted 42m links while only 0.6m appeared in next 1 year
- Prioritize predicted edges:
  - Global Threshold: pick top L links with highest weight
  - Local Threshold: at each node, pick L' edges with highest weight
  - Adaptive Local Threshold: at each node v, pick links determined by a function of its degree f(deg(v))



Predicted edge

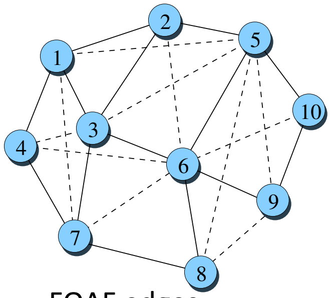


Node degree in Jan08 vs. Jan 09

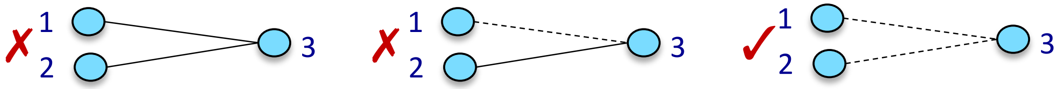
Stats:
•53% of new edges predicted by FOAF
•98% of true positives *chosen* by adaptive thresholding

Dynamic Anonymization: Our Approach

- Approach:
  - Use link prediction to predict edges $\hat{E}_t$ at time t that involve new users $U_t \setminus U_{t-1}$
  - Prediction-based Grouping: Place new users $u_1,u_2 \in U_t \setminus U_{t-1}$ into a group if $(u_1,u_2) \notin E_t$, and there is no path of length two with at most one predicted edge between u_1,u_2



- - FOAF edges

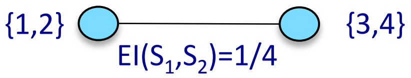


- Output: Anonymized graph $G'_t = G'_{t-1} \cup \{V_t \setminus V_{t-1}, E_t \setminus E_{t-1}\}$
- Properties:
  - Privacy: Link prediction preempts weakening of privacy due to new edges, no provable guarantee
  - Utility: accurately answer queries on any G'_t and on the sequence G'_0, G'_1, …, G'_t

Experiments: Privacy Loss

Edge Identification: EI(S_1,S_2) is the probability with which an attacker can determine the presence of an edge between a pair of users in groups S_1 and S_2

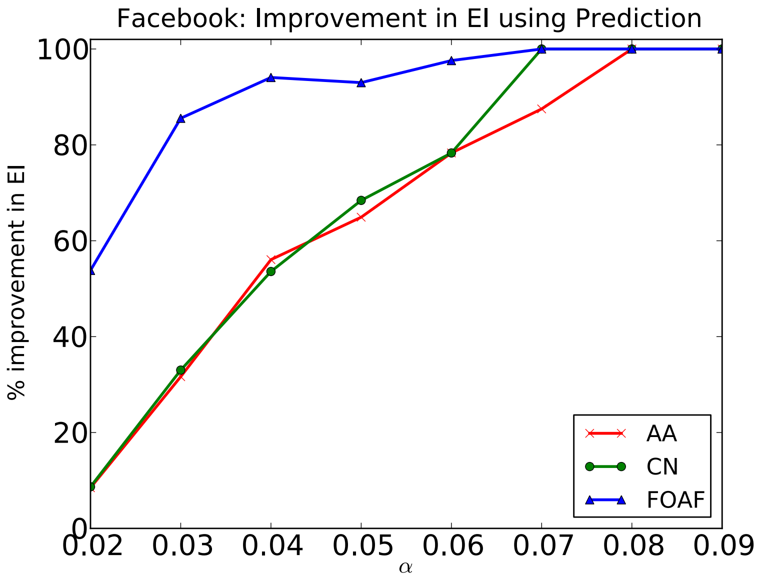
$$EI(S_1, S_2) = \frac{|(S_1 \times S_2) \cap E_t|}{|S_1||S_2|}$$



{1,2}

{3,4}

EI(S_1,S_2)=1/4

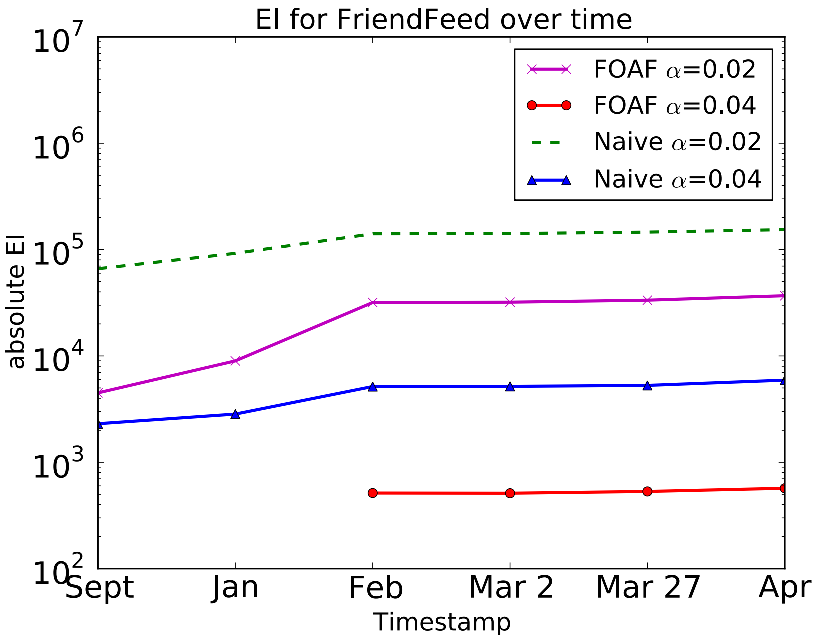


Data: 54K nodes, 887K edges in Jan 08 of New Orleans Facebook network. 630K new edges added by Jan 09.

Graph EI: number of pairs of groups S_1,S_2 with EI(S_1,S_2)≥α, for 1/k^2≤α≤1

About 90% improvement in EI with prediction-based grouping using FOAF edges, relative to Naïve approach (without prediction) for k=10

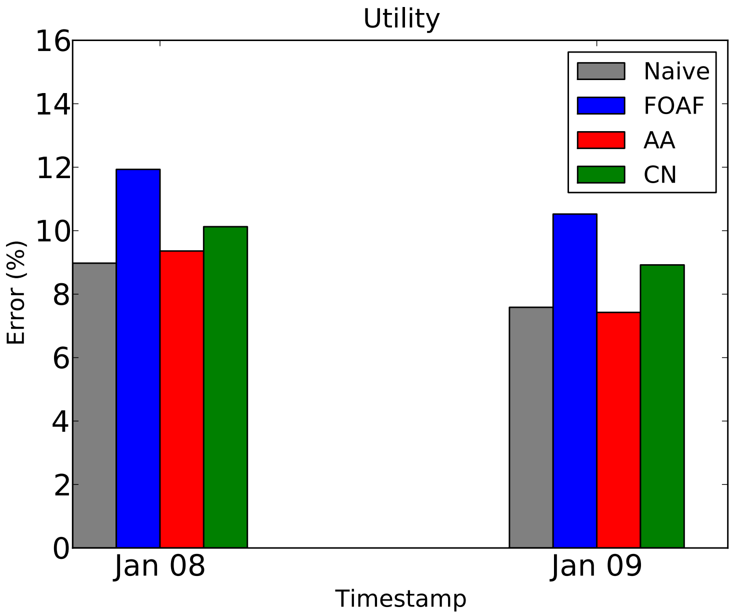
Experiments: Multiple releases



Absolute value of Graph EI is consistently higher when graph is anonymized without prediction

Friendfeed Data: 99K nodes, 1.4M edges in Sept 08.
803K new edges added in 7 months.
60% edges are predicted correctly by FOAF

Experiments: Utility



Facebook Data: New Orleans graph

Query: “How many Facebook users aged 15-20 interact with users aged 20-30 at the start and end of the measurement period?”

There is a about 3% increase in the error in query answer with FOAF anonymized graph.

Conclusions

- Presented prediction based anonymization to publish multiple instances
- Proposed new privacy metrics to evaluate loss in privacy due to multiple published instances
- Experiments show 90% improvement in privacy, with no significant loss in utility

Thank you!

And, a special thanks to: CCICADA, Rutgers for supporting part of this research