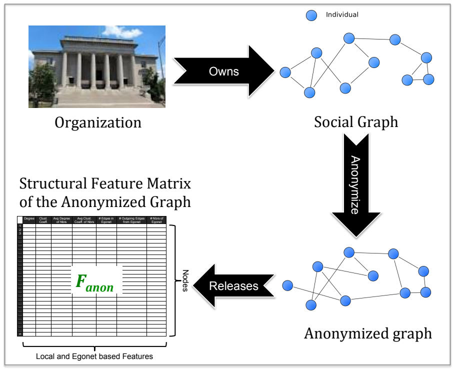
Finding the Most Appropriate Auxiliary Data for Social Graph Deanonymization

Priya Govindan*   Sucheta Soundarajan   Tina Eliassi-Rad

Problem & Motivation

How can an adversary select the most appropriate auxiliary graph to breach the privacy of individuals in an anonymized graph?



Goal of the adversary

Find an auxiliary graph G_{aux}, whose structural feature matrix (F_{aux}) has high *Overlap* and low *Lookalikes* with F_{anon}

The true node ***Overlap*** is the fraction of the nodes in F_{aux} that appear in F_{anon}

Lookalikes for a node x in F_{aux} is the number of nodes in F_{anon} that are at least as similar to x, as its matching node x' in F_{anon}

Lookalikes for a graph-pair is the average lookalikes of nodes in F_{aux}, normalized by size of F_{anon}

Challenges

1. No link structure
2. Nodes have many lookalikes (i.e., similar structural features)
3. Difficult to distinguish between nodes in F_{anon} that are present in F_{aux} and those that are absent in F_{aux}

Approach & Results

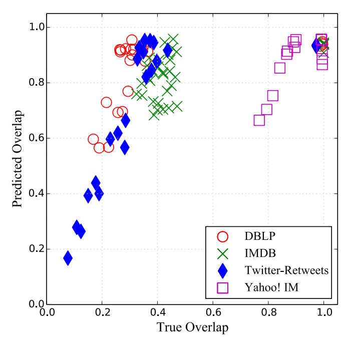
Case 1: Adversary has no side info

Given: F_{anon} and F_{aux}

Predicted ***Overlap*** = *Maximum Overlap* ×

$$\left(1 - Canberra\left(Centroid(F_{anon}), Centroid(F_{aux})\right)\right)$$

where the ***Maximum Overlap*** is the minimum of $|F_{anon}|$ and $|F_{aux}|$, divided by $|F_{aux}|$



When the predicted node-overlap is low (< 0.5), then the true node-overlap is also low (< 0.2)

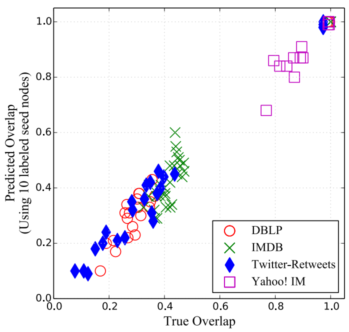
The features for each node in F_{anon} are:

1. Node's degree
2. Avg. degree of node's neighbors
3. Node's clustering coefficient
4. Avg. clustering coefficient of node's neighbors
5. # of edges between node's neighbors
6. # of nodes adjacent to node's neighbors
7. # of edges outgoing from the node's neighbors

Case 2: Adversary has labels/ matches for some nodes

Given: F_{anon}, F_{aux}, and labels (present/absent) for k nodes selected uniformly at random

Predicted ***Overlap*** = ratio of `present' labels in k seeds

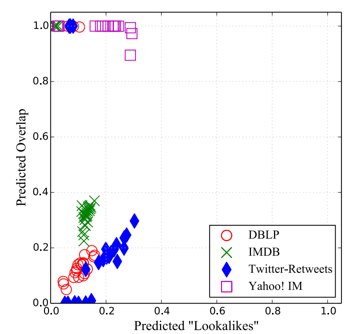


Given 10 seed-labels chosen uniformly at random, overlap can be predicted quite well (MAE = 0.03)

Given: F_{anon}, F_{aux}, matches for $m\%$ nodes

Predicted ***Overlap*** = ratio of predicted 'present' labels, by learning a classifier on labeled nodes

Predicted ***Lookalikes*** = Average lookalikes of seed matches



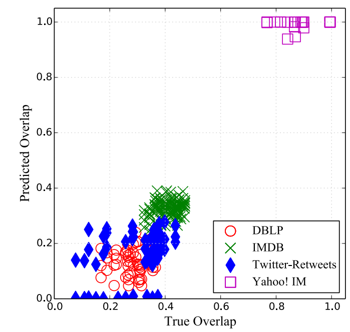
Given 10% seed-matches, overlap can be predicted very accurately (~100%) when the predicted lookalikes is < 0.1

Case 3: Adversary has labels on another auxiliary graph

Given: F_{anon}, $F1_{aux}$, $F2_{aux}$, labels on all nodes of $F1_{aux}$

A classifier is trained on the labels of $F1_{aux}$; is used to predict labels on the nodes of $F2_{aux}$

Predicted ***Overlap*** between F_{anon} and $F2_{aux}$ = ratio of predicted 'present' labels, by the classifier



Most values lie on the diagonal (with RMSE = 0.14 from the 45-degree line), thus transfer-learning predictions are deemed good estimates of the true overlap

Conclusion

- Selecting the most appropriate auxiliary data for deanonymization of F_{anon} reduces to the problem of predicting the amount of node-overlap between F_{anon} and F_{aux}
- Given **no additional info**, an adversary can identify graphs with low ***Overlap*** with F_{anon}
- Given **labels for some of the nodes**, an adversary can predict the ***Overlap*** quite well. If also given **some seed matches**, an adversary can estimate the ***Lookalikes*** for the given graph pair
- Given **labels for one graph**, an adversary can learn to predict ***Overlap*** in another graph.

* priyagn@cs.rutgers.edu