

FINDING THE MOST APPROPRIATE AUXILIARY DATA FOR SOCIAL GRAPH DEANONYMIZATION

Priya Govindan Sucheta Soundarajan Tina Eliassi-Rad <u>{priyagn, s.soundarajan, eliassi}@cs.rutgers.edu</u>









There are so many auxiliary data.



There are so many auxiliary data.



There are so many auxiliary data.



Q. Which auxiliary data should the adversary chose for her deanonymization task?

There are so many auxiliary data.



Q. Which auxiliary data should the adversary chose for her deanonymization task?

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

Overlap & Lookalikes

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

Overlap & Lookalikes

- The true node Overlap is
 - the fraction of 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}

Overlap & Lookalikes

- The true node Overlap is
 - the fraction of 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

No link structure



Challenges

- No link structure
- Nodes have many lookalikes (i.e., similar structural features)



Challenges

- No link structure
- Nodes have many lookalikes (i.e., similar structural features)

8/24/14

• Difficult to distinguish between nodes in F_{anon} that are present in F_{aux} and those that are absent in F_{aux}



We consider three cases

- **Case 1:** Adversary has no side information
 - An adversary can identify graphs that have low (< 0.5) Overlap with *F*_{anon}

8/24/14

We consider three cases

- Case 1: Adversary has no side information
 - An adversary can identify graphs that have low (< 0.5) Overlap with *F*_{anon}
- Case 2: Adversary has labels/matches for some nodes
 - With labels for a handful of randomly selected nodes, an adversary can predict the Overlap quite well (with MAE = 0.03)
 - With some (~10%) seed matches, an adversary can identify graphs that have very high (~100%) Overlap and very low (<0.1) Lookalikes with F_{anon}

8/24/14

We consider three cases

- Case 1: Adversary has no side information
 - An adversary can identify graphs that have low (< 0.5) Overlap with *F*_{anon}
- Case 2: Adversary has labels/matches for some nodes
 - With labels for a handful of randomly selected nodes, an adversary can predict the Overlap quite well (with MAE = 0.03)
 - With some (~10%) seed matches, an adversary can identify graphs that have very high (~100%) Overlap and very low (<0.1) Lookalikes with F_{anon}
- Case 3: Adversary has labels on another auxiliary graph
 - An adversary can accurately learn to predict Overlap in another graph

Thank you

RUTGERS

