De-anonymizing Social Networks

Anton Petrov
CS 457
10/11/11

Overview

- Anonymity does not equal privacy:
  - Anonymity is when your identity is hidden.
  - Privacy is having control over the access to your personal information.
  - Example: surfing with TOR vs. using SSL.
- Data sanitization only leads to anonymity.
- Availability of large datasets compromises privacy.
- Differential privacy as a possible solution.

Compromising privacy

- Corporations & government agencies do not keep data to themselves.
  - Using APIs to crawl and aggregate data.
  - Targeted advertising.
  - Third party applications.
  - Public datasets – Census, Genome information on AWS.

- Sanitization
  - Changes to dataset prior to release.
  - NULL-ing.
  - Substitution.
  - Masking of data – credit cards.

Compromised privacy examples

- Netflix 'breach' in 2007.
  - Prize of $1,000,000.
  - Cross reference data with IMDB ratings.
  - Movie ratings unique after you eliminate top 100.
  - Note: users still anonymous but their privacy was compromised in the sense that users submitted their movie ratings to Netflix believing that those ratings would remain private.
- AOL fiasco – 2006.
  - Meant for research. Once on the Internet, always on the Internet.
  - Semantic identification : Thelma Arnold.
  - User 927.
- Latanya Sweeney – Linked medical records to US Census data and managed to retrieve medical record for governor of Massachusetts.

Narayanan & Shmatikov

- Main contribution: demonstrated large scale feasibility & introduced the idea of self-reinforcing feedback.

- Social network can be modeled using a (directed) graph:
  - Entities are represented by nodes & node attributes.
  - Relationships are represented by edges & edge attributes.

- Privacy
  - Node and edge attributes.
  - Who wants to breach users' privacy?
    - Classify attackers based on their capabilities & goals
      - Government
      - Agencies & advertisers
      - Creeps

Why is an active attack unfeasible?

- Active attack = creation of dummy nodes by adversary.
  - Fundamental assumption is that adversary can modify a network prior to its release.
- Prohibitively expensive.
- Dummy 'cluster' will have no incoming nodes => raise suspicion.
- Mutual link is required for the release of node & edge attribute information. Real users are unlikely to link to dummy nodes.
- Instead focus on passive attack.

The algorithm: notation & setup

- Social network, S, is modeled using a directed graph $G = (V, E)$.
  - Set of attributes for each $v \in V$ denoted by X and similarly for edges, the set of edge attributes is denoted by Y.
- Researchers simulated a sanitized graph by picking a sub-graph of their crawled data and introducing some noise by removing edges and adding a few fake ones.
- Assumption is that adversary has access to an auxiliary network which has minimal overlap with the target network.
  - This is a *very* realistic assumption.
  - Access to an auxiliary network does not mean most of the work is already done.

The algorithm: notation & setup

- Auxiliary network information
  - Aggregate – just a regular social network with nodes & edges. This information is used in 'propagation' stage of the algorithm.
  - Individual – detailed information about a very small number of members of the target network.
    - Used in 'seed identification' stage of algorithm.
    - Adversary must be able to identify these entities in auxiliary aggregate network.
    - Not difficult to obtain this information.

- Main objective: node re-identification
  - Any subsequent privacy breach will be more effective if you have information about the end points of the edge.

The algorithm

- Seed identification – brute force approach, search target graph for sub graph that corresponds to the individual auxiliary information obtained by adversary.

- Propagation – takes as input the target and auxiliary graphs along with a seed mapping, obtained in the previous step.
    - Start with the accumulated list of mapped pairs between *V1* & *V2.*
    - Pick an arbitrary unmapped node *u* in *V1.*
    - Compute a score for each unmapped node *v* in *V2*, equal to the number of neighbors of *u* that have been mapped to neighbors of *v.*
    - If the strength of the match is above a certain threshold then add the mapping to our set. Do the nodes have the same neighbors?
    - We could consider probabilistic mappings but a deterministic one would be easier to understand.

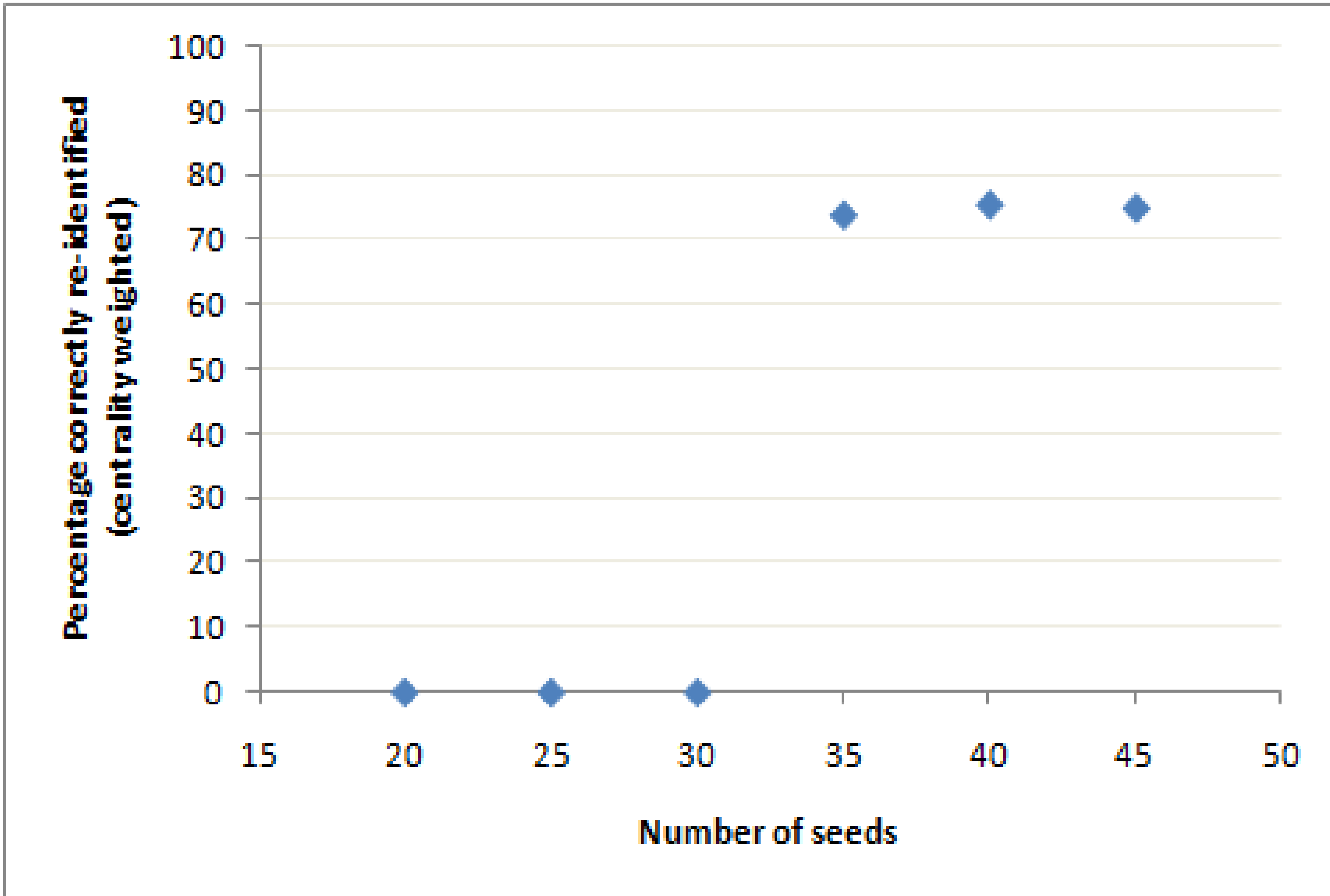
The algorithm

- Eccentricity
  - How much does an item X stand out from the rest?
  - *[max (X) – max2 (X)] / σ (X)*
- Edge directionality?
  - Given that our graph is directed, we first compute score for incoming edges, then score for outgoing edges and then sum.
- Node degree?
  - The mapping scores will be biased in favor of nodes with a high degree. To compensate we divide score by square root of node degree.
- Measuring success
  - Do not use fraction of nodes identified – singular node problem.
  - Instead use the concept of node centrality.
    - Measure importance of node using its degree.

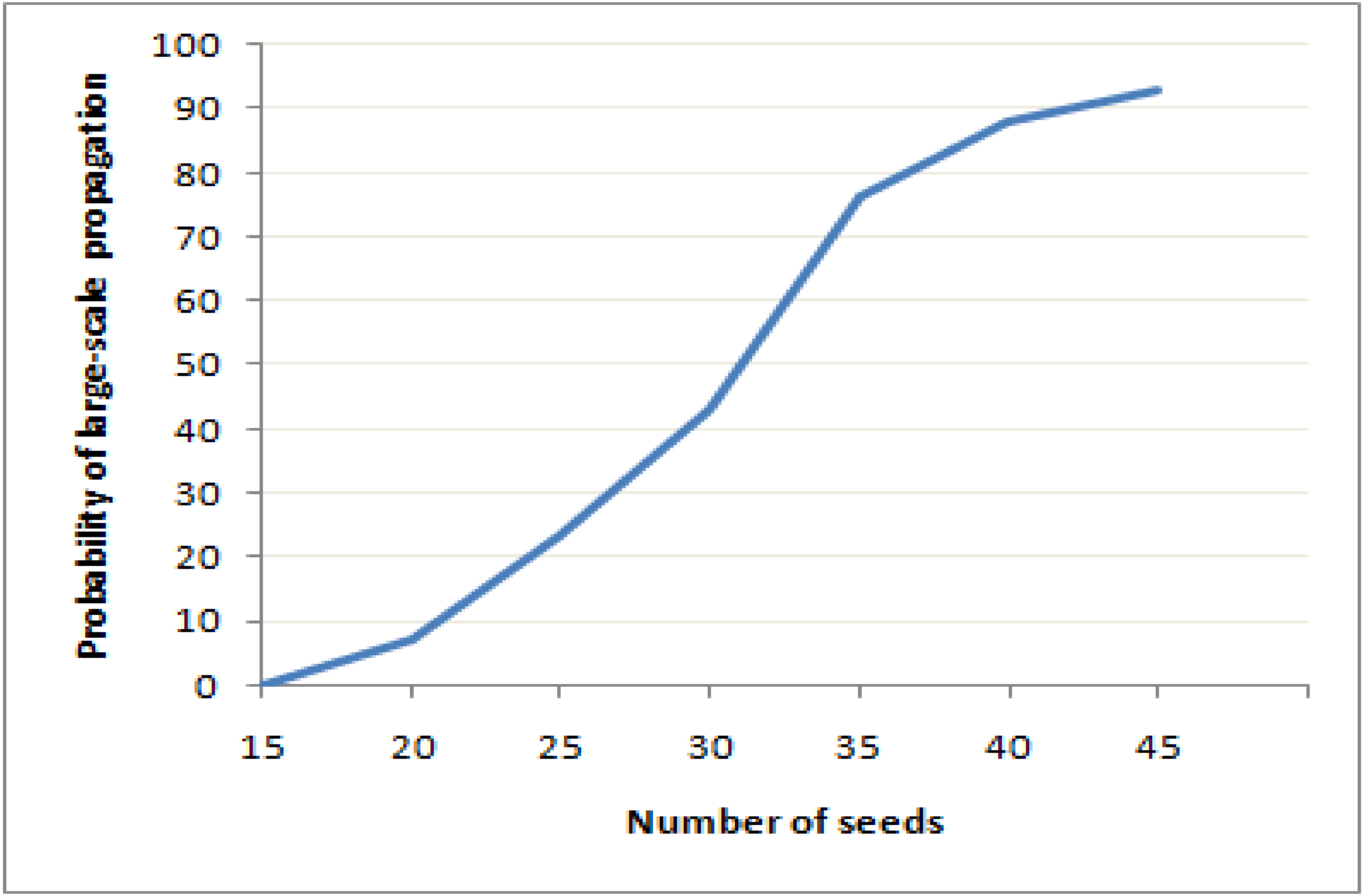
Data & Performance

All measurements were done with a node overlap of 25% and an edge overlap of 50%.

Data & Performance

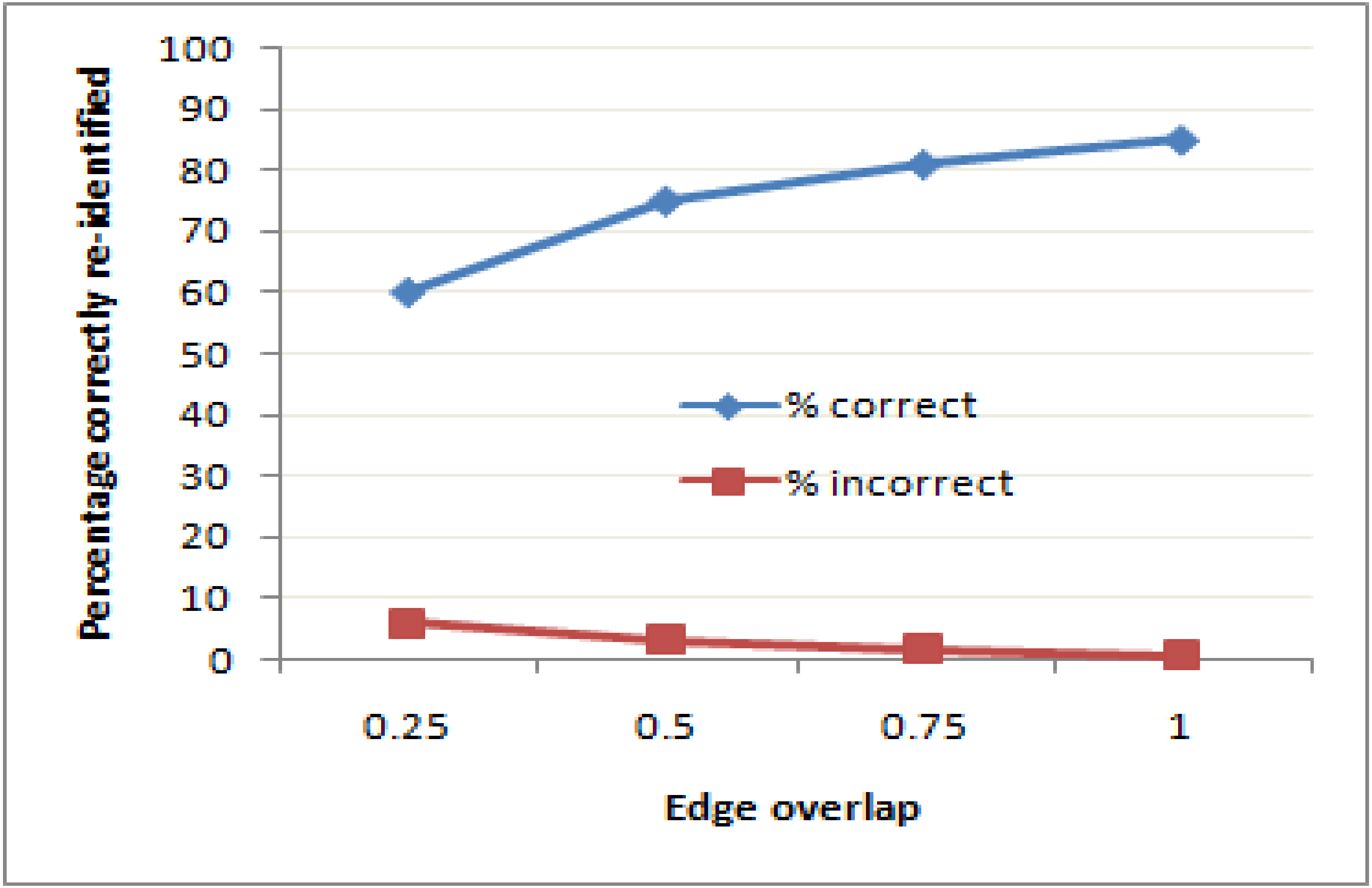


Data & Performance



Self-reinforcement & feedback is crucial and needs a substantial initial seed.

Data & Performance



Effect of noise. Node overlap 25%. Number of seeds 50.

[My] Solution

- Use the concept of *differential privacy* for graphs.

- Initially defined for statistical databases only.
  - Introduced by Cynthia Dwork.
  - Aggregate data into a database – US Census is a great example.
  - Allow people to query that database and extract information in such a way that no individual record can be inferred.

- In a perfect world we would have an equivalent of semantic security for databases.
  - Impossible because an adversary will have auxiliary information.
  - Instead think of privacy as being differential:
    - Your participation in a database should not significantly increase the chance of you being exposed.

Differential Privacy

- Interactive vs. non-interactive
  - Curator sits between database & users.
  - Curator computes and publishes some statistics.

- Numerical definition
  - $Pr[K(D1) \in S] \leq \exp(e) \times Pr[K(D2) \in S]$
  - D1 and D2 are data sets that differ by one element
  - K () is the randomizing function, S is Range (K())

- Sensitivity of some query f()
  - $\Delta f = \max \{D1, D2\} ||f(D1) - f(D2)||$
  - How great a difference should be hidden by the noise
  - $K(X) = f(X) + (Lap(\Delta f / e))$

Differential Privacy for Graphs

- *A Differentially Private Graph Estimator,* Mir & Wright.

- Develop method of generating a synthetic graph that will give users a fairly accurate picture of the graph while preserving the privacy of individuals.

- Assuming that observed data is generated from an underlying (unknown) distribution, the paper suggests a technique of using the observed data to produce an estimator for the underlying distribution.

- Graphs can then be sampled from this distribution and hopefully they will have similar properties to the original.

Quick Conclusions

- Lots of data.
- Lots of unsecured data that anyone can mine.
- Corporations & government agencies need to improve their data sanitization by starting to think about differential privacy.
- This problem will not go away – data will keep growing.

Ze Questions

Bibliography

N. Anderson. *Pulling back the curtain on "anonymous" Twitterers.* http://arstechnica.com/tech-policy/news/2009/03/pulling-back-the-curtain-on-anonymous-twitterers.ars/. Published: 03-31- 09. Accessed: 10-09-11.

N. Anderson. *"Anonymized" data really isn't – and here's why not.* http://arstechnica.com/tech-policy/news/2009/09/your-secrets-live-online-in-databases-of-ruin.ars/. Published: 09-08-09. Accessed: 10-09-11.

M. Darakhshan and R. Wright. *A Differentially Private Graph Estimator.* IEEE International Conference on Data Mining Workshops (2009), 122-129.

C. Dwork. *Differential Privacy*. Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (invited), 2006

C. Dwork. *Differential Privacy: A Survey of Results*. Proceedings of The 5th Annual Conference on Theory and Applications of Models of Computation (invited), 2008

A. Narayanan and V. Shmatikov. *De-anonymizing Social Network.* http://www.cs.utexas.edu/~shmat/. Accessed: 10-09-11.

B. Schneier. *Why 'Anonymous' Data Sometimes Isn't.* http://www.wired.com/politics/security/commentary/securitymatters/2007/12/securitymatters_1213/ Published: 12-13-07. Accessed: 10-09-11.