De-anonymizing Social Networks

Arvind Narayanan and Vitaly Shmatikov

Presented by:
Chance Shafor
Nicole Shiver
Benjamin Woodruff
Aaron Marquez
Monetization: Targeted Advertising

“… breakthrough technology that uses social graph data to dramatically improve online marketing … "Social Engagement Data" consists of anonymous information regarding the relationships between people”

The critical distinction … between the use of personal information for advertisements in personally-identifiable form, and the use, dissemination, or sharing of information with advertisers in non-personally-identifiable form.
Third-Party Applications

Race to platform domination

“WidgetLaboratory was gathering credentials from users and otherwise creating havoc on Ning networks”

“A security hole exposed the birthdays and relationship status of strangers, including Facebook executives, [and] the wife of Google co-founder Larry Page”
Call graphs

2 trillion edges

Examples of outsourced call graphs

<table>
<thead>
<tr>
<th>Country</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hungary</td>
<td>2.5M</td>
</tr>
<tr>
<td>France(?)</td>
<td>7M</td>
</tr>
<tr>
<td>India</td>
<td>3M</td>
</tr>
</tbody>
</table>

3,000 companies providing wireless services in the U.S
“Jefferson High”: Romantic and Sexual Network

Real data!
What needs protecting?

- Node attributes
  - Name, phone no.
  - Sexual orientation
- Edge attributes
  - Date of creation
  - Strength
- Edge existence
Contribution 1: De-anonymization Attack

- Large-scale, robust against noise
- Based on mapping between two different networks

Previous work: [Narayanan & Shmatikov 08]
Anonymity vs. Privacy

- Anonymity is **insufficient** for privacy
- Anonymity is **necessary** for privacy
- Anonymity is **unachievable** in practice
Prior Work: Active Attacks

- Needs thousands of “sybil” nodes
- Easy for network operators to identify and block
- Doesn’t work well for mutual-connection networks

See: Backstrom et al.
Contribution 2: Privacy Framework

• Model for data
• Model for data release, “sanitization”
• Background knowledge (Auxiliary information)
• What does privacy breach mean?
  – What needs to be protected?
  – What is the attack model?
• How does anonymity relate to privacy?
Prior Work: Defenses

- Some strategies ignore aux information, limited attacker models
- Zero-knowledge system limitations
- k-anonymity doesn’t work
Background Knowledge

Aux = Auxiliary information = Background knowledge

Edges don’t necessarily overlap (even within the overlapping nodes)

Aux is large!
Aux is a graph!
Aux is very imprecise

[BDK07]: Aux is not large-scale
Contribution 3:
Experimental Validation

Previous work (claiming positive results)
- Simulated data lacks good models
- Too small, too sparse

This work
- Real data
- Millions of nodes
- High average degree
- Across datasets
De-anonymization attack → privacy breach

Large-scale Background Knowledge

Attack model
Social Network Model

- Social network ‘S’ consisting of graph $G = (V, E)$
- Attributes $X$ for each node in $V$
- Attributes $Y$ for each node in $E$
- Discrete attributes are values in $\{0, 1\}$
- Released subsets $V_{san}, E_{san}, X_{san},$ and $Y_{san}$ as “anonymized” graph of $S_{san}$
Motivating scenario: Overlapping networks

• Social networks A and B have overlapping memberships
  – The topology of B is public

• Owner of A shares anonymized graph along with sensitive attributes
  – say, to enable targeted advertising

• Can advertiser learn sensitive information from released graph A’?
Attacker Model

- In addition to $S_{san}$, attacker has overlapping $S_{aux}$
- It may be possible to extract $S_{aux}$ from $S_{aux}$
- Nation-state actors will already have $S_{aux}$
Attackers

Global surveillance

Phishing

Spammer
Abusive advertising/marketing

Nosy friend
Why is the problem hard?

Recall: no names, identifiers; just topology

So.. graph isomorphism?

Harder! (Noise)

Key insights:

two stage paradigm
self-reinforcing matching
Re-identification: Two-stage Paradigm

• Seed identification:
  – Detailed knowledge about local graph structure
  – Small number of nodes
  – Relatively precise
  – Used to create initial “seed” mapping

• Propagation:
  – Imprecise knowledge of large number of nodes
  – “infects” large proportion of nodes
What is a Breach?

**Ground Truth:** A mapping $\mu_G$ between $V_{aux}$ and $V_{san}$

**Re-Identification:** Finding a probabilistic $\mu$

$$\mu: V_{san} \times (V_{aux} \cup \{\bot\}) \rightarrow [0,1]$$

where

$$\mu(v_{aux}, v_{san})$$

is the probability that $v_{aux}$ maps to $v_{san}$.
Measuring success

\[ V_{mapped} = \{ v \in V_{aux} : \mu_G(v) \neq \perp \} \]

**Success rate:**
\[
\frac{\sum_{v \in V_{mapped}} \text{PR}[\mu(v) = \mu_G(v)]}{\sum_{v \in V_{mapped}} \nu(v)}
\]

**Error rate:**
\[
\frac{\sum_{v \in V_{mapped}} \text{PR}[\mu(v) \neq \perp \land \mu(v) \neq \mu_G(v)]}{\sum_{v \in V_{mapped}} \nu(v)}
\]
Seed Identification: Background Knowledge

How:
- Creating sybil nodes
- Bribing
- Phishing
- Hacked machines
- Stolen cellphones

What: List of neighbors
- Degree
- Number of common neighbors of two nodes

Recall: only a small number of seeds (a few dozens) needed for attack to work
Seed identification – results

Even with 10% noise, a clique of 4 users (out of 5.5 million) can be reidentified with 65% probability

Re-identifying clique of 4 users in anonymized graph (Livejournal, 5.5M nodes)
Two-stage Paradigm

Target
- De-identified nodes
- Contains sensitive attributes

“Seed” mapping

Aux
- Identifiable nodes
- No sensitive information
Propagation: Scoring

Much harder on real networks
15% edge overlap!
Propagation: Heuristics

- Edge directionality
- Edge weights
- Node weights
- Mapping strength
  - Eccentricity
  - Mapping size
  - Cosine
- Reverse map
  - Imperfect match
- Non-bijective
- Deletion
- Parameter rotation
- Self-reinforcing matching
Eccentricity

- It measures how much an item in a set $X$ "stands out" from the rest

\[
(\text{max}(X) - \text{max}_2(X)) / \text{stddev}(X)
\]

- Where $\text{max}$ and $\text{max}_2$ denote the highest and second highest values, respectively
function propagationStep(lgraph, rgraph, mapping)
    for lnode in lgraph.nodes:
        scores[lnode] = matchScores(lgraph, rgraph, mapping, lnode)
        if eccentricity(scores[lnode]) < theta: continue
        rnode = (pick node from rgraph.nodes where
            scores[lnode][node] = max(scores[lnode]))
        scores[rnode] = matchScores(rgraph, lgraph, invert(mapping), rnode)
        if eccentricity(scores[rnode]) < theta: continue
        reverse_match = (pick node from lgraph.nodes where
            scores[rnode][node] = max(scores[rnode]))
        if reverse_match != lnode:
            continue
        mapping[lnode] = rnode
function matchScores(lgraph, rgraph, mapping, lnode)
    // cosine similarity

function eccentricity(items)
    return (max(items) - max2(items)) / std_dev(items)

until convergence do:
    propagationStep(lgraph, rgraph, seed_mapping)
Perturbation Strategy

- Make two copies of $E$ and independently delete edges at random from each copy.
- Project the copies on $V_1$ and $V_2$, respectively to get $E_1$ and $E_2$.

Optimal link prediction!
Whenever virality occurs, percentage re-identified is remarkably constant

Prob. of achieving virality is strongly dependent on number of seeds
Ground Truth

Datasets: flickr twitter

Compute $\mu_G$ using username, name, and location

• Longer usernames, rarer or longer names are better matches
• Combine these attributes to compute a score
• Reject scores below a threshold
Mapping Two Large Networks

15% edge overlap

Throw away identifying info
Start from **160 seeds**, apply propagation algo.
**31% of common users re-identified**
  Further 12% incorrectly identified
  (weighted by centrality)

How do we know?
  “Ground truth” oracle based on identifying info.
Summary

Anonymization doesn’t work on social graphs
   – Regardless of style of privacy definition
   – Adding noise doesn’t help

Solutions
   – PII should disappear from privacy laws
   – Opt-in rather than opt-out
   – Reputable advertising as a business model
Questions?
Extra slides
Anonymity is necessary for privacy
Anonymity does not seem achievable in real-world social networks
33bits.org

Chronicling the end of data anonymity
Similarity Measure

• Captures how much information we are recovering about a record

• Threshold function on individual attributes
  – For the domain of each attribute, boolean function says when two values are “similar”
  – Different measures for different attributes

• Cosine similarity
  \[ \text{Sim}(r_1, r_2) = \frac{\sum \text{Sim}(r_{1i}, r_{2i})}{|\text{supp}(r_1) \cup \text{supp}(r_2)|} \]
Typical behavior of propagation
Social Networks: Data Release

• Select subset of nodes
• Compute induced subgraph
• Sanitize edges
• Select attributes
• Publish
function bestMatch(left_node, mapping)
{
    initialize scores with zeros

    for left_nbr in node.nbrs()
    {
        right_nbr = mapping[left_nbr]
        for right_nbr in right_nbr.nbrs():
            scores[right_node] += 1
    }

    return right_node with highest score
}

Deceptively simple!
Differential privacy [move to end]

• Remove one user, database looks the same
• Social network: local changes do have global effects
  – Tipping point/phase transition
• This is the kind of info. we are often interested in
• Removing user can disconnect graph
• Enforcement: int. mechanisms, real numbers, relational db’s
• Graph invariant properties: GCC, max clique, cut size
Data Collection and Sharing

2010 census: $15B

Talk to sociologists!

Most significant turning point in the history of sociology
Defining privacy breach via re-identification

What sensitive information can the adversary learn (by any means whatsoever)?

What fraction of nodes can the adversary re-identify?

What sensitive information can the adversary learn (via node re-identification?)
Anonymity != Privacy

Anonymity is insufficient for privacy

Clusters around (sensitive) attributes

- Detecting terrorist activity based on clusters
  - Hayes 07, Soghoian 08

- Dating circles for disease positive individuals

- Interest communities
  - “Identifying dense subgraphs in large graphs”
Anonymity != Privacy

Anonymity is insufficient for privacy

Released dataset

No "identifiable person"

Background knowledge
The Crutch of Anonymity

Anonymity is **unachievable in practice**

[Narayanan & Shmatikov 08]

**The critical distinction** ... between the use of personal information for advertisements in personally-identifiable form, and the use, dissemination, or sharing of information with advertisers in **non-personally-identifiable form**.
Active vs. Passive Attack

Limitations
- Only OSN’s
- Can be detected, sanitized
- Small scale

Passive attack feasible as long as background knowledge is available

Backstrom, Dwork, Kleinberg [WWW07]

Attacker controlled node

Nice try cold-calling people
Depends on users “adding you back”
Tens of nodes compromised
The Crutch of Anonymity

[Narayanan & Shmatikov 08]
What doesn’t work

Anonymity is **insufficient** for privacy

<table>
<thead>
<tr>
<th>Authors</th>
<th>Conference/Report</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hay et. al.</td>
<td>Tech report, 2007</td>
<td>Degree sequences, approximate nbhd’s</td>
</tr>
<tr>
<td>Lui and Terzi</td>
<td>SIGMOD 08</td>
<td>Only node degrees</td>
</tr>
<tr>
<td>Campan and Truta</td>
<td>PinKDD 08</td>
<td>Attributes as well as structure</td>
</tr>
<tr>
<td>Zhou and Pei</td>
<td>ICDE 08</td>
<td>1-neighborhoods</td>
</tr>
</tbody>
</table>

Syntactic techniques
Other Examples

Doctor-patient network

Faces in social context

Pseudonymous users
Aggregation

Universal data accessibility

Aggregation services that exploit data portability