Attacking Anonymized Social Network

From: Wherefore Art Thou RX3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography

Presented By:

Machigar Ongtang  (Ongtang@cse.psu.edu)
Social Networks

- A social structure made of nodes representing individuals or organization, linked by some interdependency (*friendship, trade, communication, web links, etc.*)

- Some Statistics:
  - MySpace -- 206,304,468 user accounts
  - Windows Live (MSN) Spaces -- 120,000,000 user accounts
  - Orkut (by Google) -- 67,962,551 user accounts
  - Hi5 -- 50,000,000 user accounts
  - Friendster -- 50,000,000 user accounts
  - Facebook -- 48,000,000 user accounts
  - LiveJournal -- 12,900,000 user accounts
Research on Social Networks

- **Use of Public Data:** publicly crawlable blogging, social networking sites, etc.  \(\rightarrow\) _no privacy concern_

- **Use of Sensitive Data:** Email, Instant Messaging, closed online communities, private network  \(\rightarrow\) _need privacy protection!!_

**Pure Social Network Data**

![Diagram of social network data](attachment:image.png)

- Alice has sent messages to Bob
- Alice includes Bob in her address book
Anonymized Social Networks

- **Privacy**: Protect privacy of individual users.
- **Utilities**: Preserving global network properties to study network structure, dynamics, clustering patterns, etc.
- Replace the true user’s name with a random user ID.

Is this enough? Can anonymization protect users’ privacy?
Attacks on Anonymized Networks

- **Auxiliary Information**: Analysis of content, time series, and logs
- **WITHOUT Auxiliary Information**: Consider Link/ No Link ***

Basic security ...

- **Privacy Breach**: Identify nodes and learn the edge relations among them.
- **Passive Attack**: Observe the released anonymized social network without interfering.
  - Undetectable
- **Active Attack**: Create some new nodes (e.g. new email accounts). Create (patterned) edges among new nodes and to victim nodes.
  - Hard to detect.
Before releasing the anonymized network $G$ of $n-k$ nodes, attacker:

- Choose a set of $b$ targeted users.
- Create a subgraph $H$ containing $k$ nodes.
- Attach $H$ to the targeted nodes.

Creating the subgraph $H$ --> **structural steganography**
Active Attack

After the anonymized network is release:

- Find the subgraph H in the graph G
- Follow edges from H to locate b target nodes and their true location in G
- Determine all edges among these b nodes --> breach privacy
Active Attack -- Finding subgraph H

- Subgraph H must be uniquely & efficiently identifiable regardless of G.
- No other subgraph $S \neq H$ in G s.t. $G[S]$ and H are isomorphic.
- Subgraph H has no automorphism.

Graph Isomorphism

- A one-to-one mapping between vertices of two graph P and Q.
- Isomorphism $f: P \rightarrow Q$ Two vertices $u$ and $v$ in P are connected if their corresponding node $f(u)$ and $f(v)$ are connected in Q.
- Automorphism = isomorphism to itself

$P$ and $Q$ are isomorphic
Walk-Based Attack

- Randomly generate subgraph \( H = (x_1, x_2, \ldots, x_k) \) with \( k = \Theta(\log n) \)
- Link each targeted node \( w_i \) to distinct subset of nodes in \( H \)
- Create each edge within \( H \) with probability of 0.5
- Number of compromised nodes \( b = \Theta(\log^2 n) \)

**Recall:** We need \( H \) that
- **Efficiently identifiable:** Efficiently findable in unlabeled graph
- **Unique in term of isomorphism:** If a subgraph with the same structure as \( H \) is found, it is actually \( H \).
- **No automorphism:** When \( H \) is found, we know which node is which and can correctly label \((x_1, x_2, \ldots, x_k)\)
Walk-Based Attack -- Construct H

- $H$ = set of nodes X size $k = (2+\delta) \log n$ ($\delta > 0$)
- $W$ = set of targeted users size $b = O(\log^2 n)$
  - e.g. $n = 1000M$, $b = 900$, $k \approx 30$
- External degree for node $x_i$: $\Delta_i \in [d_0, d_1]$ for $d_0 \leq d_1 = O(\log n)$
- Each $w_j$ connects to a set of nodes $N_j \subseteq X$. Set
- $N_j$ must be of size at most $c=3$ and are distinct across all nodes $w_j$. 

![Diagram showing target nodes and H set]
Walk-Based Attack -- Construct H

- Add arbitrary edges from H to G-H to make it $\Delta_i$ for all $x_i$.
- Add internal edges in H: edge $(x_i, x_{i+1})$
- Add additional internal edges connecting $(x_i, x_j)$ with probability 0.5
- Therefore, each node $x_i$ has total degrees of $\Delta'_i = \Delta_i + (#internal \ edges)$
Walk-Based Attack -- Finding H

- **Degree Test**: Node $x_i$ has total degrees of $\Delta_i' = \Delta_i + (\# \text{ internal edges})$
- **Internal Structure Test**: Node $x_i$ links to correct subset of $\{x_1, x_2, \ldots, x_{i-1}\}$
- **Search tree $T$**: All nodes $\alpha_i$ in $T$ has corresponding node $f(\alpha_i)$ in $G$.
- Every path of nodes $\alpha_1, \alpha_2, \ldots, \alpha_j$ from the root must have corresponding path in $G$ formed by nodes $f(\alpha_1), f(\alpha_2), \ldots, f(\alpha_j)$ with the same degree sequence $x_1, x_2, \ldots, x_j$.
- The probability of a false path surviving to depth $l \approx 2^{-l^2/2}$
Walk-Based Attack -- Uniqueness of H

**F0:** With high probability, there is no subset of nodes $S \neq X$ in $G$ such that $G[S]$ is isomorphic to $G[X] = H$

**Non-overlapping Case:** $S$ disjoint from $X$

**Graph H:** $k$ nodes

$$\binom{k}{2} \approx \frac{k^2}{2} \quad \text{Possible edges} \quad \rightarrow \quad 2^{\frac{k^2}{2}} \quad \text{Possible graphs}$$

**Subgraph G-H:** select $k$ nodes from $n$

$$\binom{n}{k} < n^k \approx 2^{k \log n} \quad \text{Possible subgraphs}$$

**Probability of isomorphic:**

$$P = \frac{2^{k \log n}}{2^{\frac{k^2}{2}}} \quad \rightarrow \quad \text{Drop quickly when } k > 2\log(n)$$

**Example:** $n=12M$

Choose $k = 2\log(12M) = 14 \rightarrow \quad P \approx \frac{2^{99}}{2^{99}} = 1$

Choose $k = (2+\delta) \log n = 15 \rightarrow \quad P \approx \frac{2^{106}}{2^{113}} = 0.011$
Walk-Based Attack -- Uniqueness of H

- Overlapping Case -- \( G[S] \text{ and } G[X] \) is isomorphic with \( S \) overlaps \( X \)

\[
P \approx \sum_{j \geq 1} k^{c_2 \log k} \left( \frac{2^{3.5} k^2}{n^\delta} \right)^j
\]

Drop quickly as \( n \) increases and \( k > 2 \log(n) \)
Walk-Based Attack -- Uniqueness of H

**F1:** For $c_1 > 4$, there is no disjoint sets of nodes $Y$ and $Z$ in $H$, each of size $c_1 \log(k)$, such that $H[Y]$ and $H[Z]$ are isomorphic.

- Scope down what we had from F0:
  - Graph $G$ size $n$ $\Rightarrow$ Subgraph $H$ size $k$.
  - Sets of nodes size $(2 + \delta) \log n$ $\Rightarrow$ Sets of nodes size $(c_1 > 4) \log(k)$.

---

**Fixed Point of a Isomorphism**

For Isomorphism mapping $S$ to $S'$ ($f: S \rightarrow S'$):

- A fixed point is in both $S$ and $S'$.
- A fixed point maps to itself.
Claim 3.1: Let $A$, $B$ and $Y$ are disjoint sets of nodes in $G$ with $B, Y \subseteq X$. With isomorphism $f: A \cup Y \rightarrow B \cup Y$, $|\{ f(A) \not\in B \}| \leq c_1 \log(k)$ nodes.

- Consider path from $A \rightarrow Y \rightarrow B$: $|Y'|$ and $|B'|$ are disjoint and $H[B']$ and $H[Y']$ are isomorphic. $\Rightarrow |Y'| \leq c_1 \log(k)$ $\Rightarrow$ # paths $A \rightarrow Y \rightarrow B \leq c_1 \log(k)$

- $Y' = f(A) \not\in B$ $\Rightarrow |f(A) \not\in B| \leq c_1 \log(k)$
Claim 3.2: Let $A$, $B$ and $Y$ be disjoint sets of nodes in $G$ with $B,Y \subseteq X$. With isomorphism $f: A \cup Y \rightarrow B \cup Y$, set $Y$ has at most $c_2 \log(k)$ nodes that are not fixed point of $f$, where $c_2 \geq 3c_1$.

Choose every other edge in the path or cycle. In cycle, choose 1 edges from 3 z-nodes. $\Rightarrow$ Worst case $|Z|/3$ edges

- $Z_1$ and $Z_2$ are disjoint subset of $X$; and $G[Z_1]$ and $G[Z_2]$ are isomorphic
- $\Rightarrow$ From F1, $|Z_1| = |Z_2| \leq c_1 \log(k)$
- $|Z_1| = |Z_2| = \#selected\ edges = |Z|/3 \Rightarrow |Z|/3 \leq c_1 \log(k)$
- $|Z| \leq 3c_1 \log(k) \leq c_2 \log(k)$
Walk-Based Attack -- Experiment

- Use network of relationship links on blogging of LiveJournal (4.4M nodes and 77M edges); Anonymized it
- Investigate the ranges parameter for successful attack.
Cut-Based Attack

- Theoretical asymptotic lower bound for #new nodes: $\Omega(\sqrt{\log n})$
- Randomly generate subgraph $H = (x_1, x_2, \ldots, x_k)$ with $k = O(\sqrt{\log n})$
- Number of compromised nodes $b = \Theta(\sqrt{\log n})$

Construction of $H$

- For $W=(w_1, w_2, \ldots, w_b)$ $b$ targeted users, create $X= (x_1, x_2, \ldots, x_k)$ where $k = 3b+3$ nodes
- Create links between each pair $(x_i, x_j)$ with probability = 0.5
- Choose arbitrary $b$ nodes $(x_1, x_2, \ldots, x_b)$; connect $x_i$ to $w_i$
Cut-Based Attack -- Construct H

- $d(H) = \text{min degree in } H$
- $c(H) = \text{min internal cut in } H$

**Properties:**

With high probability

- $b = \text{size of cut between } H \text{ and } G-H$
- $c(H) = d(H) \geq k/3 > b$
- $H$ has non-trivial automorphism

**Observe**

- All internal cuts in $H > b$
- Cuts of size $\leq b$ are external cuts between $H$ and $G-H$. They will never break $H$. 
Recall: Cuts of size $\leq b$ are external cuts between $H$ and $G-H$. They will never break $H$.

- **Step1:**
  - Use Gomory-Hu tree to break the graph along the cuts of size $\leq b$
  - Finally, one of these chunks is $H$

- **Step2:**
  - Find which one is $H$
  - $H$ needs to be unique
Cut-Based Attack -- Uniqueness of $H$

Graph $H$: $k$ nodes

\[
\binom{k}{2} \approx \frac{k^2}{2} \quad \text{Possible edges} \quad \rightarrow \quad 2^{k^2/2} \quad \text{Possible graphs}
\]

Subgraph $G-H$:
- There are $n/k$ sets.
- Each set has $k!$ possible graphs

\[
(n/k)k! \quad \text{Possible subgraphs}
\]

Probability of isomorphic:

\[
P = \frac{(n/k)k!}{2^{k^2/2}} \quad \rightarrow \quad \text{Drop quickly when } k > \sqrt{\log(n)}
\]

Example: \( n=1000 \text{M} \)

Log(n) = 9

Choose $k = 12 \rightarrow P = 8.45271119 \times 10^{-6}$
Gomory-Hu Tree

- Tree with the same set of nodes in G. Edge are labeled with weight.

- The value of min-cut for (u,v)
  
  = #edges on the smallest cut that will disconnect u and v
  
  = min-weight on the path between u and v in T

- Breaking graph G along the cuts of size \( \leq b \)
  
  = delete all edges of size \( \leq b \) from T

- Repeat until all forests have size \( k \)

- Brute force to check whether each forest is isomorphic to H
Walk-Based VS Cut-Based Attacks

<table>
<thead>
<tr>
<th>Walk-Based Attack</th>
<th>Cut-Based Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ Fast recovery algorithm</td>
<td>▪ More expensive recovery algorithm</td>
</tr>
<tr>
<td>▪ Hard to detect</td>
<td>▪ Easier to detect because $H$ is dense and tends to stand out</td>
</tr>
<tr>
<td>▪ Need more new nodes $\Theta(\log n)$</td>
<td>▪ Need less new node $O(\sqrt{\log n})$ (close to theoretical asymptotic lower bound: $\Omega(\sqrt{\log n})$)</td>
</tr>
<tr>
<td>▪ Can detect $b = \Theta(\log^2 n) = \Theta(k^2)$</td>
<td>▪ Can detect only $\Theta(\sqrt{\log n}) = \Theta(k)$</td>
</tr>
</tbody>
</table>
Passive Attack

- **Community of Interest**: most nodes in social network usually belong to a small uniquely identifiable subgraph.
- An attacker can collude with other k-1 friends to identify additional nodes connected to the distinct subset of the coalition.

**Assumptions**
- All colluders know edges among themselves, i.e. internal structure of H.
- All colluders know the name of their neighbors outside the coalition.
- There may be no Hamiltonian Path linking $x_1 \rightarrow x_2 \rightarrow \ldots \rightarrow x_k$
Passive Attack -- Finding H

Search Tree T

- Degree Test
- For ALL subset $S \subseteq \{1, \ldots, k\}$, node $\alpha$ matching H must have $g_\alpha(S) = g(S)$
  - If we consider $S = \{1, 3, 5\}$
  - $g(S) = q$: There is $q$ users that connects to $x_1$, $x_3$, and $x_5$. 
Passive Attack

![Graphs showing probability and number of users compromised with different algorithms and coalition sizes.](image-url)
Active VS Passive Attacks

<table>
<thead>
<tr>
<th>Active Attack</th>
<th>Passive Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ More effective. Work with high probability in any network.</td>
<td>▪ Attackers may not be able to identify themselves after seeing the released anonymized network.</td>
</tr>
<tr>
<td>▪ Can choose the victims</td>
<td>▪ The victims are only those linked to the attackers.</td>
</tr>
<tr>
<td>▪ Risk of being detected</td>
<td>▪ Harder to detect</td>
</tr>
</tbody>
</table>

Semi-Passive Attack:

▪ Create only additional links to the targeted nodes. No additional node.
▪ Can breach privacy on the scale approaching that of the active attack.
Possible Countermeasures

- Random Perturbation
  - m-perturbation
  - Randomly delete m edges and insert m edges

- Model-based Perturbation
  - Derive statistical model from original data
  - Develop model to bias the perturbation to give desired properties of the graph
  - Give better utility
Conclusion

- Anonymized network is not safe.
- Already published anonymized networks are susceptible to passive attack, which does not require advanced planning.