Privacy Risk in Anonymized Heterogeneous Information Networks
(How to Break Anonymity of the KDD Cup 2012 Dataset)

Aston Zhang¹, Xing Xie², Kevin C.-C. Chang¹
Carl A. Gunter¹, Jiawei Han¹, Xiaofeng Wang³

University of Illinois at Urbana-Champaign¹
Microsoft Research²
Indiana University at Bloomington³
K-Anonymity

- Any tuple can be re-identified with a probability no higher than 1/K
  - AAA BBB, 3-anonymity
  - AABBBB, 2-anonymity
  - ABBB, 1-anonymity
K-Anonymity?

• If the adversary (attacker) is only interested in breaking privacy of tuples with value B...
  – AAA BBBB, 3-anonymity
  – ABBBBB, 4-anonymity?
  – ABBBBBB, 5-anonymity?
Limitations of K-Anonymity

$T_{1000}$ (1000 tuples of the same value: AAA...AA): 1000-anonymity
$T_2$ (500 pairs of same values: AABBCCDD...): 2-anonymity

Consider injecting a globally unique tuple $t^*$

$T_{1000}^*$ (1000 same tuples: AAAAAA...AAAA $t^*$)
$T_2^*$ (500 pairs of same values: AABBCCDD...$t^*$)

• Are their security level really the same in terms of k-anonymity?
K-Anonymity: Worst Says All
Different Individuals May Have Different Privacy Needs
Our Proposed Privacy Risk

• Individual Privacy Risk – *Allow us to differentiate!*

• Aggregate *Individual Privacy Risk* to obtain Dataset Privacy Risk

  – Mathematical Factor
  
  – Social Factor
• Privacy risk of tuple $t_i$ in dataset $T$:

$$\mathcal{R}(t_i) = \frac{l(t_i)}{k(t_i)}$$

Individual Privacy Risk

- $l(t_i)$: loss function (social component)
- $k(t_i)$: the number of tuples in $T$ with the same value of $t_i$ (math component)

• Privacy risk of dataset $T$:

$$\mathcal{R}(T) = \frac{\sum_{i=1}^{N} \mathcal{R}(t_i)}{N}$$

Dataset Privacy Risk

- $N$: the size of dataset $T$ (the number of tuples in $T$)
Lemma 1: Given dataset $T$ with cardinality $C(T)$, for each tuple $t_i$ in $T$, assuming the loss function is independent of $1/k(t_i)$ with mean value $\mu$, the expected privacy risk

$$E(R(T)) = \frac{\mu C(T)}{N}$$
• **Theorem 1:** The privacy risk $R(T)$ of dataset $T$ is

$$R(T) = \frac{C(T)}{N}, \quad (R(T) \in \left[ \frac{1}{N}, 1 \right])$$

- $N$: the number of tuples in $T$
- $C(T)$: cardinality of $T$ —— number of distinct (combined) attribute values describing tuples

  - $T=ABCDE$: $C(T)=5$, $R(T)=1$
  - $T'=AAAAA$: $C(T')=1$, $R(T')=1/5$
  - $T''=AAA$: $C(T'')=1$, $R(T'')=1/3$
Privacy Risk = \( \frac{C(T)}{N} \), Better Interpretation

\( T_{1000} \) (1000 same tuples: AAAAAA...AAAA)
- 1000-anonymity (Privacy Risk: 1/1000)

\( T_2 \) (500 pairs of same values: AABBCCDD...)
- 2-anonymity (Privacy Risk: 500/1000)

Inject globally unique tuple \( t^* \)

\( T_{1000}^* \) (1000 same tuples: AAAAAA...AAAA\( t^* \))
- 1-anonymity (Privacy Risk: 2/1001)

\( T_2^* \) (500 pairs of same values: AABBCCDD...\( t^* \))
- 1-anonymity (Privacy Risk: 501/1001)

\( R(T_{1000}^*) < R(T_2^*) \)
How Bad Is Our Privacy in HIN?

- Heterogeneous Information Networks (HIN)
We Live In a More Connected World

• Heterogeneous Information Networks (HIN)
  – Multiple types of entities (nodes) or links (edges)
  – Present in
    • social media (Twitter)
    • medical information systems (EMR)
    • academic information systems (DBLP)
    • and more...
Heterogeneous Information Networks in Social Media (KDD Cup 2012 Dataset)
Network Schema
Meta-Paths

- **user mention path**: User \( \xrightarrow{\text{post}} \) Tweet \( \xrightarrow{\text{mention}} \) User or User \( \xrightarrow{\text{post}} \) Comment \( \xrightarrow{\text{mention}} \) User (short-circuited feature: mention strength)

- **user retweet path**: User \( \xrightarrow{\text{post}} \) Tweet \( \xrightarrow{\text{retweet}} \) Tweet \( \xrightarrow{\text{posted by}} \) User (short-circuited feature: retweet strength)

- **user comment path**: User \( \xrightarrow{\text{post}} \) Comment \( \xrightarrow{\text{comment}} \) Tweet \( \xrightarrow{\text{posted by}} \) User or User \( \xrightarrow{\text{post}} \) Comment \( \xrightarrow{\text{comment}} \) Comment \( \xrightarrow{\text{posted by}} \) User (short-circuited feature: comment strength)

- **user follow path**: User \( \xrightarrow{\text{follow}} \) User
The neighbors of the target entity A1X are generated along target meta paths
• **Theorem 2**: For power-law distribution of the user out-degree, the lower and upper bounds for the expected heterogeneous information network cardinality grows faster than double exponentially with respect to the max. distance of utilized neighbors.

Privacy Risk = $C(T)/N$
Privacy Risk Increases With More Link Types

Table 1: Privacy Risk of the Anonymized t.qq Dataset (density: 0.01, size: 1000) increases as the amount of utilized target network schema link types increases (in percentage)

<table>
<thead>
<tr>
<th>Types of Links</th>
<th>Max. Distance</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td></td>
<td>84.4</td>
<td>93.8</td>
<td>93.8</td>
</tr>
<tr>
<td>m</td>
<td></td>
<td>85.4</td>
<td>93.6</td>
<td>93.8</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td>87.6</td>
<td>93.6</td>
<td>93.9</td>
</tr>
<tr>
<td>r</td>
<td></td>
<td>90.2</td>
<td>94.2</td>
<td>94.3</td>
</tr>
<tr>
<td>f-m</td>
<td></td>
<td>96.0</td>
<td>98.5</td>
<td>98.6</td>
</tr>
<tr>
<td>f-c</td>
<td></td>
<td>95.6</td>
<td>98.5</td>
<td>98.5</td>
</tr>
<tr>
<td>f-r</td>
<td></td>
<td>96.8</td>
<td>98.5</td>
<td>98.5</td>
</tr>
<tr>
<td>m-c</td>
<td></td>
<td>89.9</td>
<td>94.0</td>
<td>94.2</td>
</tr>
<tr>
<td>m-r</td>
<td></td>
<td>91.2</td>
<td>94.4</td>
<td>94.5</td>
</tr>
<tr>
<td>c-r</td>
<td></td>
<td>91.8</td>
<td>94.4</td>
<td>94.5</td>
</tr>
<tr>
<td>f-m-c</td>
<td></td>
<td>96.5</td>
<td>98.5</td>
<td>98.6</td>
</tr>
<tr>
<td>f-m-r</td>
<td></td>
<td>96.9</td>
<td>98.6</td>
<td>98.6</td>
</tr>
<tr>
<td>f-c-r</td>
<td></td>
<td>96.8</td>
<td>98.6</td>
<td>98.6</td>
</tr>
<tr>
<td>m-c-r</td>
<td></td>
<td>92.3</td>
<td>94.5</td>
<td>94.6</td>
</tr>
<tr>
<td>f-m-c-r</td>
<td></td>
<td>96.9</td>
<td>98.6</td>
<td>98.6</td>
</tr>
</tbody>
</table>

*f: follow; m: mention; r: retweet; c: comment

*Max. Distance n: max. distance of utilized neighbors to target entities

*n = 0: only target entities’ profiles are utilized and risk is always 1.1%
Privacy Risk Increases With More Link Types
DeHIN Algorithm to Prey on Privacy Risk in HIN

Algorithm 1 De-anonymizing entity $v'$ in Heterogeneous Information Networks: DeHIN ($G$, $G'$, $T_G^*$, $v'$, $l$)

**Inputs:** $G = (V, E)$: public graph, $G' = (V', E')$: private graph, $T_G^* = (E^*, L^*)$: target network schema, $v' \in G'$: target entity, $l$: specified utilized level of neighborhoods

**Outputs:** $\mathcal{C}$: candidate set from public data that match $v'$

1. $\mathcal{C} \leftarrow \emptyset$
2. for all $v \in V$ IN $G = (V, E)$ do
   3. if entity_attribute_match($v'$, $v$, $E^*$) then
      4. if $l > 0$ then
         5. if link_match($l$, $v'$, $v$, $G$, $G'$, $T_G^*$) then
            6. $\mathcal{C} \leftarrow \mathcal{C} \cup \{v\}$
         7. end if
      8. else
        9. $\mathcal{C} \leftarrow \mathcal{C} \cup \{v\}$
      10. end if
   11. end if
3. end for
4. return $\mathcal{C}$
DeHIN Algorithm to Prey on Privacy Risk in HIN

Algorithm 2 Comparing neighborhoods of two entities \( v' \) and \( v \) via heterogeneous links: \( \text{link\_match}(l, v', v, G, G', T) \)

Inputs: 
- \( l \): level of neighborhoods utilized, \( v' \in G' \): target entity, \( v \): the entity in public graph under comparison,
- \( G = (V, E) \): public graph, \( G' = (V', E') \): private graph,
- \( T = (\mathcal{E}, \mathcal{C}) \): target network schema

Outputs: \( \text{is\_match} \): a boolean value

\[
\text{is\_match} \leftarrow \text{true}
\]
\[
G_B \leftarrow \emptyset \quad \text{(The bipartite graph modeling neighborhood matching)}
\]
\[
\mathcal{N}_v(v', \mathcal{L}_t) \leftarrow v'\text{'s neighborhoods via the link type } \mathcal{L}_t
\]
\[
\mathcal{N}_v(v, \mathcal{L}_t) \leftarrow v\text{'s neighborhoods via the link type } \mathcal{L}_t
\]
for all link type \( \mathcal{L}_i \) IN \( \mathcal{L}^* \) do
  for all neighbor \( n'_i \) IN \( \mathcal{N}_v(v', \mathcal{L}_t) \) do
    \( \emptyset \leftarrow C(n'_i) \) \quad (\( C(n'_i) \): candidate set for \( n'_i \))
    for all neighbor \( n_i \) IN \( \mathcal{N}_v(v, \mathcal{L}_t) \) do
      if \( \text{link\_attribute\_match}(n'_i, n_i) \) then
        if \( \text{entity\_attribute\_match}(n'_i, n_i) \) then
          if \( l = 1 \) then
            \( C(n'_i) \leftarrow n'_i \)
          else
            if \( \text{link\_match}(l - 1, v', v, G, G', T) \) then
              \( C(n'_i) \leftarrow n'_i \)
            end if
          end if
        end if
      end if
    end for
  end for
end for
\[
G_B \leftarrow C(n'_i)
\]
end for
if \( \text{max\_bipartite\_match}(G_B) \neq |\mathcal{N}_v(v', \mathcal{L}_t)| \) then
  \( \text{is\_match} \leftarrow \text{false} \)
end if
return \( \text{is\_match} \)
Break Anonymity of the KDD Cup 2012 Dataset

Advantages: Exploit the identified privacy risk without requiring creating new accounts or relying on easily-detectable graph structures in a large-scale network [BDK’07, NV’09]
Experiments – The Existing Defense

Modified based on [Wu and Ying’11]
Experiments – Key Measures

\[
Precision = \frac{\sum_{i=1}^{V'} s(v'_i)}{|V'|}
\]

\[
Reduction \ Rate^8 = \frac{1}{|V'|} \sum_{i=1}^{V'} \left(1 - \frac{|C(v'_i)|}{|V|}\right)
\]

\[
density = \frac{|E|}{m|V|^2 + (|\mathcal{L}| - m)|V|(|V| - 1)}
\]
DeHIN Performs Better on Denser Networks by Utilizing Longer-Distance Neighbors

<table>
<thead>
<tr>
<th>Density</th>
<th>Max. Distance 0</th>
<th></th>
<th>Max. Distance 1</th>
<th></th>
<th>Max. Distance 2</th>
<th></th>
<th>Max. Distance 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Reduction Rate</td>
<td></td>
<td>Precision</td>
<td>Reduction Rate</td>
<td></td>
<td>Precision</td>
<td>Reduction Rate</td>
</tr>
<tr>
<td>0.001</td>
<td>4.1</td>
<td>99.836</td>
<td></td>
<td>12.6</td>
<td>99.848</td>
<td></td>
<td>12.6</td>
<td>99.848</td>
</tr>
<tr>
<td>0.002</td>
<td>5.1</td>
<td>99.925</td>
<td></td>
<td>22</td>
<td>99.947</td>
<td></td>
<td>22.7</td>
<td>99.948</td>
</tr>
<tr>
<td>0.003</td>
<td>6.5</td>
<td>99.917</td>
<td></td>
<td>32.8</td>
<td>99.944</td>
<td></td>
<td>33.5</td>
<td>99.945</td>
</tr>
<tr>
<td>0.004</td>
<td>4.3</td>
<td>99.907</td>
<td></td>
<td>39.4</td>
<td>99.941</td>
<td></td>
<td>40.8</td>
<td>99.942</td>
</tr>
<tr>
<td>0.005</td>
<td>4.3</td>
<td>99.927</td>
<td></td>
<td>48.7</td>
<td>99.969</td>
<td></td>
<td>49.8</td>
<td>99.969</td>
</tr>
<tr>
<td>0.006</td>
<td>7</td>
<td>99.920</td>
<td></td>
<td>59.4</td>
<td>99.979</td>
<td></td>
<td>61.6</td>
<td>99.980</td>
</tr>
<tr>
<td>0.007</td>
<td>5.1</td>
<td>99.908</td>
<td></td>
<td>65.6</td>
<td>99.977</td>
<td></td>
<td>68.8</td>
<td>99.978</td>
</tr>
<tr>
<td>0.008</td>
<td>5.3</td>
<td>99.921</td>
<td></td>
<td>76.6</td>
<td>99.989</td>
<td></td>
<td>78.8</td>
<td>99.989</td>
</tr>
<tr>
<td>0.009</td>
<td>6.4</td>
<td>99.914</td>
<td></td>
<td>86.2</td>
<td>99.997</td>
<td></td>
<td>88.6</td>
<td>99.997</td>
</tr>
<tr>
<td>0.01</td>
<td>5.4</td>
<td>99.892</td>
<td></td>
<td>92.5</td>
<td>99.989</td>
<td></td>
<td>95.6</td>
<td>99.990</td>
</tr>
</tbody>
</table>

*Max. Distance n: max. distance of utilized neighbors to target entities; when n = 0, only target entities’ profile attributes are utilized.
### DeHIN Performs Better with More Link Types

Table 3: Performance of DeHIN on t.qq anonymized dataset (density: 0.01) improves as the amount of utilized target network schema link types increases (in percentage)

<table>
<thead>
<tr>
<th>Types of Links</th>
<th>Max. Distance 1</th>
<th></th>
<th>Max. Distance 2</th>
<th></th>
<th>Max. Distance 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Reduction Rate</td>
<td>Precision</td>
<td>Reduction Rate</td>
<td>Precision</td>
<td>Reduction Rate</td>
</tr>
<tr>
<td>f</td>
<td>68.1</td>
<td>99.982</td>
<td>77.6</td>
<td>99.983</td>
<td>77.7</td>
<td>99.983</td>
</tr>
<tr>
<td>m</td>
<td>80.9</td>
<td>99.976</td>
<td>87.8</td>
<td>99.976</td>
<td>88</td>
<td>99.976</td>
</tr>
<tr>
<td>c</td>
<td>82.8</td>
<td>99.975</td>
<td>88.7</td>
<td>99.976</td>
<td>88.8</td>
<td>99.976</td>
</tr>
<tr>
<td>r</td>
<td>81.1</td>
<td>99.976</td>
<td>88.7</td>
<td>99.976</td>
<td>88.9</td>
<td>99.976</td>
</tr>
<tr>
<td>f-m</td>
<td>89.3</td>
<td>99.989</td>
<td>94.2</td>
<td>99.990</td>
<td>94.2</td>
<td>99.990</td>
</tr>
<tr>
<td>f-c</td>
<td>90.1</td>
<td>99.989</td>
<td>94.6</td>
<td>99.990</td>
<td>94.6</td>
<td>99.990</td>
</tr>
<tr>
<td>f-r</td>
<td>89.2</td>
<td>99.989</td>
<td>94.9</td>
<td>99.990</td>
<td>95</td>
<td>99.990</td>
</tr>
<tr>
<td>m-c</td>
<td>84.7</td>
<td>99.976</td>
<td>89.6</td>
<td>99.976</td>
<td>89.7</td>
<td>99.976</td>
</tr>
<tr>
<td>m-r</td>
<td>83.2</td>
<td>99.976</td>
<td>89.5</td>
<td>99.977</td>
<td>89.7</td>
<td>99.977</td>
</tr>
<tr>
<td>c-r</td>
<td>85.2</td>
<td>99.976</td>
<td>90.3</td>
<td>99.976</td>
<td>90.5</td>
<td>99.976</td>
</tr>
<tr>
<td>f-m-c</td>
<td>91.6</td>
<td>99.989</td>
<td>94.8</td>
<td>99.990</td>
<td>94.8</td>
<td>99.990</td>
</tr>
<tr>
<td>f-m-r</td>
<td>90.6</td>
<td>99.989</td>
<td>95.1</td>
<td>99.990</td>
<td>95.2</td>
<td>99.990</td>
</tr>
<tr>
<td>f-c-r</td>
<td>91.5</td>
<td>99.989</td>
<td>95.4</td>
<td>99.990</td>
<td>95.5</td>
<td>99.990</td>
</tr>
<tr>
<td>m-c-r</td>
<td>86.5</td>
<td>99.977</td>
<td>91</td>
<td>99.977</td>
<td>91.2</td>
<td>99.977</td>
</tr>
<tr>
<td>f-m-c-r</td>
<td>92.5</td>
<td>99.989</td>
<td>95.6</td>
<td>99.990</td>
<td>95.7</td>
<td>99.990</td>
</tr>
</tbody>
</table>

* f: follow; m: mention; r: retweet; c: comment

*Max. Distance n: max. distance of utilized neighbors to target entities; when \( n = 0 \), only target entities’ profile attributes are utilized

*\( n = 0 \): only target entities’ profiles are utilized—precision and reduction rate are always 5.4% and 99.892%
DeHIN’s Performance Slightly Degrades When Anonymity Gets Stronger

(a) Density: 0.001  
(b) Density: 0.002  
(c) Density: 0.003  
(d) Density: 0.004  
(e) Density: 0.005  
(f) Density: 0.006  
(g) Density: 0.007  
(h) Density: 0.008  
(i) Density: 0.009  
(j) Density: 0.01
DeHIN Performs Better with More Link Types

![Graph showing precision with varying DeHIN’s Max. Distance of Utilized Neighbors n](image)
Privacy Risk Increases With More Link Types
The Research Roadmap

Utilized Heterogeneity Information

Utilized Attribute Information

DeHIN

Utilized Graph Information

- k-closeness
- l-diversity
- k-anonymity
- k-degree
- k-neighbors
- k-symmetry
- k-automorphism
- k-security