SYBILFUSE: Combining Local Attributes with Global Structure to Perform Robust Sybil Detection

Peng Gao \(^1\) Binghui Wang \(^2\) Neil Zhenqiang Gong \(^2\)
Sanjeev R. Kulkarni \(^1\) Kurt Thomas \(^3\) Prateek Mittal \(^1\)

\(^1\)Princeton University
\(^2\)Iowa State University
\(^3\)Google
1. Introduction to Sybil Attack

2. Background and Related Work

3. The SYBILFUSE Framework

4. Evaluation on Labeled Twitter Networks

5. Conclusion
Sybil Attack: A single adversary injects multiple colluding identities in the system to compromise security and privacy.
Sybil Attack: A single adversary injects multiple colluding identities in the system to compromise security and privacy.

83 million Facebook accounts are fakes and dupes

By Heather Kelly, CNN
Updated 5:27 AM ET, Fri August 3, 2012
**Sybil Attack**: A single adversary injects multiple colluding identities in the system to compromise security and privacy.

83 million Facebook accounts are fakes and dupes

By Heather Kelly, CNN

Updated 5:27 AM ET, Fri August 3, 2012

Rise of the Twitter bots: Social network admits 23 MILLION of its users tweet automatically without human input

- Twitter now has more than 270 million users who actively log in and tweet
- Of these active users, approximately 23 million never visit the site
- Instead, they pull information from Twitter automatically using other apps
Sybil Attack: Impact

- Malware
- Fake reviews
- Spam messages
- Fake news
- Scams
- Private data
- Unsolicited friend requests
- Others
- Others

Sybil Attack
Sybil Attack: Network Model

Benign Region

Sybil Region

Attack Edges
Outline

1 Introduction to Sybil Attack

2 Background and Related Work

3 The SYBILFUSE Framework

4 Evaluation on Labeled Twitter Networks

5 Conclusion
Local Attributes-Based Approaches

- Blacklisting [Ramachandran et al. CCS’07]
- Whitelisting [Yardi et al. Firsty Monday Vol15(1)’10]
- Local structural features [Yang et al. IMC’11]
Local Attributes-Based Approaches

- Blacklisting [Ramachandran et al. CCS’07]
- Whitelisting [Yardi et al. Firsy Monday Vol15(1)’10]
- Local structural features [Yang et al. IMC’11]

Limitations:

- Sybils can mimic the behaviors of benign users by manipulating their profiles and connections.
Global Structure-Based Approaches

- SybilGuard [Yu et al. SIGCOMM’06]
- SybilLimit [Yu et al. IEEE S&P’08]
- SybilInfer [Danezis et al. NDSS’09]
- SybilRank [Cao et al. NSDI’12]
- CIA [Yang et al. WWW’12]
- SybilBelief [Gong et al. TIFS’13]
- Íntegro [Boshmaf et al. NDSS’15]
- SybilSCAR [Wang et al. INFOCOM’17]
Global Structure-Based Approaches

- SybilGuard [Yu et al. SIGCOMM’06]
- SybilLimit [Yu et al. IEEE S&P’08]
- SybilInfer [Danezis et al. NDSS’09]
- SybilRank [Cao et al. NSDI’12]
- CIA [Yang et al. WWW’12]
- SybilBelief [Gong et al. TIFS’13]
- Íntegro [Boshmaf et al. NDSS’15]
- SybilSCAR [Wang et al. INFOCOM’17]

Limitations:
- Strong-trust assumptions: limited number of attack edges
Global Structure-Based Approaches

- SybilGuard [Yu et al. SIGCOMM’06]
- SybilLimit [Yu et al. IEEE S&P’08]
- SybilInfer [Danezis et al. NDSS’09]
- SybilRank [Cao et al. NSDI’12]
- CIA [Yang et al. WWW’12]
- SybilBelief [Gong et al. TIFS’13]
- Íntegro [Boshmaf et al. NDSS’15]
- SybilSCAR [Wang et al. INFOCOM’17]

Limitations:

- Strong-trust assumptions: limited number of attack edges
  - RenRen network does not follow [Yang et al. IMC’11]
  - Link farming on Twitter [Ghosh et al. WWW’12]
Global Structure-Based Approaches

- SybilGuard [Yu et al. SIGCOMM’06]
- SybilLimit [Yu et al. IEEE S&P’08]
- SybilInfer [Danezis et al. NDSS’09]
- SybilRank [Cao et al. NSDI’12]
- CIA [Yang et al. WWW’12]
- SybilBelief [Gong et al. TIFS’13]
- Íntegro [Boshmaf et al. NDSS’15]
- SybilSCAR [Wang et al. INFOCOM’17]

Limitations:

- Strong-trust assumptions: limited number of attack edges
  - RenRen network does not follow [Yang et al. IMC’11]
  - Link farming on Twitter [Ghosh et al. WWW’12]
- Íntegro requires the number of victims to be small and the victims are accurately predicted.
1 Introduction to Sybil Attack

2 Background and Related Work

3 The SYBILFUSE Framework

4 Evaluation on Labeled Twitter Networks

5 Conclusion
SybilFuse Framework

Input
Social Network Data

Local Attributes
- Structural Attributes
- Content Attributes

Global Structure
- Directed/Undirected Graph

Known Labels

Local Classifiers
local trust scores

Trust Score Propagation
- Weighted Random Walk
- Weighted Loopy Belief Propagation

Output
- Predicted Labels
- Node Ranking
Local Trust Score Computation

$S_v$ for node $v$: probability that $v$ is benign

- Computed via training a node classifier using local node attributes (e.g., degree, local clustering coefficient, profile info)
- Normalize to $[0.1, 0.9]$
Local Trust Score Computation

$S_v$ for node $v$: probability that $v$ is benign

- Computed via training a node classifier using local node attributes (e.g., degree, local clustering coefficient, profile info)
- Normalize to $[0.1, 0.9]$ 

$S_{u,v}$ for edge $(u, v)$: probability that $u$ and $v$ take the same label (i.e., models homophily strength)

- Computed via training an edge classifier
- Similarity between node $u$ and node $v$
- Normalize to $[0.1, 0.9]$
Set the initial score of every node $v$:

$$S^{(0)}(v) = \begin{cases} 
0.9 & v \text{ is a training benign node} \\ 
0.1 & v \text{ is a training Sybil node} \\ 
S_v & \text{else}
\end{cases}$$
Set the initial score of every node $v$:

$$S^{(0)}(v) = \begin{cases} 
0.9 & \text{if } v \text{ is a training benign node} \\
0.1 & \text{if } v \text{ is a training Sybil node} \\
S_v & \text{otherwise}
\end{cases}$$

Score update equation:

$$S^{(i)}(v) = \sum_{(u,v) \in E} S^{(i-1)}(u) \frac{S_{u,v}}{\sum_{(u,w) \in E} S_{u,w}}$$
Set the initial score of every node $v$:

$$S^{(0)}(v) = \begin{cases} 
0.9 & \text{if } v \text{ is a training benign node} \\
0.1 & \text{if } v \text{ is a training Sybil node} \\
S_v & \text{otherwise}
\end{cases}$$

Score update equation:

$$S^{(i)}(v) = \sum_{(u,v) \in E} S^{(i-1)}(u) \frac{S_{u,v}}{\sum_{(u,w) \in E} S_{u,w}}$$

After $d = O(\log n)$ iterations, we obtain the final score $S^F_v$:

$$S^F_v = S^{(d)}(v)$$
Trust Score Propagation: Weighted LBP

Node & edge potentials: \( X_v \in \{1, -1\} \) represents the label of node \( v \)

\[
\psi_v(X_v) = \begin{cases} 
S_v & \text{if } X_v = 1 \\
1 - S_v & \text{if } X_v = -1 
\end{cases}
\]

\[
\psi_{u,v}(X_u, X_v) = \begin{cases} 
S_{u,v} & \text{if } X_u X_v = 1 \\
1 - S_{u,v} & \text{if } X_u X_v = -1 
\end{cases}
\]

\((G, \Psi)\) defines a pairwise Markov Random Field.
Belief update equation:

\[
m_{u \rightarrow v}(X_v) = \sum_{X_u} \left( \psi_u(X_u) \psi_{u,v}(X_u, X_v) \prod_{s \in \text{Neighbors}(u) \setminus v} m_{s \rightarrow u}(X_s) \right)
\]
Belief update equation:

\[ m_{u \rightarrow v}(X_v) = \sum_{X_u} \left( \psi_u(X_u) \psi_{u,v}(X_u, X_v) \prod_{s \in \text{Neighbors}(u) \setminus v} m_{s \rightarrow u}(X_s) \right) \]

After \( d = 5 \sim 10 \) iterations, we obtain the final score \( S^F_v \):

\[ \text{bel}_v(X_v = x_v) \propto \psi_v(X_v = x_v) \prod_{u \in \text{Neighbors}(v)} m_{u \rightarrow v}(X_v = x_v) \]

\[ S^F_v = \frac{\text{bel}_v(X_v = 1)}{\text{bel}_v(X_v = 1) + \text{bel}_v(X_v = -1)} \]
Label $L_v$ of node $v$ is predicted as:

$$L_v = \text{sign}(S_v^F - \text{threshold})$$

We can also rank nodes according to $S_v^F$. Sybil nodes with low scores will be ranked upfront.
1. Introduction to Sybil Attack

2. Background and Related Work

3. The SYBILFUSE Framework

4. Evaluation on Labeled Twitter Networks

5. Conclusion
Small Twitter Network: Measurement

- 8,167 nodes (7,358 benign nodes & 809 Sybil nodes) and 54,146 edges (40,001 attack edges)
Small Twitter Network: Measurement

- 8,167 nodes (7,358 benign nodes & 809 Sybil nodes) and 54,146 edges (40,001 attack edges)

We have the following observations:
- More than half (53.4%) of Sybils are isolated, i.e., only connect to benign nodes.
8,167 nodes (7,358 benign nodes & 809 Sybil nodes) and 54,146 edges (40,001 attack edges)

We have the following observations:

- More than half (53.4%) of Sybils are isolated, i.e., only connect to benign nodes.
- The number of attack edges is large, with 49 attack edges on average per Sybil.
8,167 nodes (7,358 benign nodes & 809 Sybil nodes) and 54,146 edges (40,001 attack edges)

We have the following observations:

- More than half (53.4%) of Sybils are isolated, i.e., only connect to benign nodes.
- The number of attack edges is large, with 49 attack edges on average per Sybil.
- More than 75% of benign nodes are victims.
8,167 nodes (7,358 benign nodes & 809 Sybil nodes) and 54,146 edges (40,001 attack edges)

We have the following observations:

- More than half (53.4%) of Sybils are isolated, i.e., only connect to benign nodes.
- The number of attack edges is large, with 49 attack edges on average per Sybil.
- More than 75% of benign nodes are victims.

Thus, the benign region and the Sybil region can hardly be viewed as separate communities.
• Incoming requests accepted ratio: $\text{Req}_{\text{in}}(v) = \frac{|\text{In}(v) \cap \text{Out}(v)|}{|\text{In}(v)|}$

• Outgoing requests accepted ratio: $\text{Req}_{\text{out}}(v) = \frac{|\text{In}(v) \cap \text{Out}(v)|}{|\text{Out}(v)|}$

• Local clustering coefficient: $\text{CC}(v) = \frac{|\{(i,j): i,j \in \text{Nei}(v), (i,j) \in E\}|}{|\text{Nei}(v)|(|\text{Nei}(v)|-1)}$
Small Twitter Network: Local Node Trust Scores

- Incoming requests accepted ratio: $\text{Req}_{\text{in}}(v) = \frac{|\text{In}(v) \cap \text{Out}(v)|}{|\text{In}(v)|}$
- Outgoing requests accepted ratio: $\text{Req}_{\text{out}}(v) = \frac{|\text{In}(v) \cap \text{Out}(v)|}{|\text{Out}(v)|}$
- Local clustering coefficient: $\text{CC}(v) = \frac{\{|(i,j): i, j \in \text{Nei}(v), (i,j) \in E\}|}{|\text{Nei}(v)|(|\text{Nei}(v)|-1)}$

We randomly sample 50 benign nodes and 50 Sybil nodes as the training set, and train a SVM classifier with RBF kernel using $\text{LIBSVM}$. 
(a) Random walk-based approaches

(b) LBP-based approaches and ensemble methods
Large Twitter Network: Measurement

- 21,297,772 nodes and 265,025,545 edges (18,414,469 attack edges)
  - 145,156 (0.7%) suspended nodes
  - 1,911,482 (9.0%) deleted nodes
  - The rest were active
Large Twitter Network: Measurement

- 21,297,772 nodes and 265,025,545 edges (18,414,469 attack edges)
  - 145,156 (0.7%) suspended nodes
  - 1,911,482 (9.0%) deleted nodes
  - The rest were active

We have the following observations:

- Half of Sybils are isolated.
- The number of attack edges is large (127 attack edges on average per Sybil).
We use the same set of features: $\text{Req}_{in}(v)$, $\text{Req}_{out}(v)$, $\text{CC}(v)$.

(a) Scatter plot

(b) CDF
Large Twitter Network: Node Feature Distribution

We use the same set of features: \( \text{Req}_{\text{in}}(v), \text{Req}_{\text{out}}(v), CC(v) \).

![Scatter plot](image1.png)

(a) Scatter plot

![CDF](image2.png)

(b) CDF

We randomly sample 3000 benign nodes and 3000 Sybil nodes as the training set, and train a SVM classifier with RBF kernel using \textit{LIBSVM}.
### AUC

<table>
<thead>
<tr>
<th></th>
<th>SR</th>
<th>CIA</th>
<th>INT</th>
<th>INT-PF</th>
<th>SB</th>
<th>SS</th>
<th>SF-RW</th>
<th>SF-LBP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.57</td>
<td>0.80</td>
<td>0.48</td>
<td>0.54</td>
<td>0.74</td>
<td>0.74</td>
<td>0.81</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Large Twitter Network: Evaluation

AUC

<table>
<thead>
<tr>
<th></th>
<th>SR</th>
<th>CIA</th>
<th>INT</th>
<th>INT-PF</th>
<th>SB</th>
<th>SS</th>
<th>SF-RW</th>
<th>SF-LBP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.57</td>
<td>0.80</td>
<td>0.48</td>
<td>0.54</td>
<td>0.74</td>
<td>0.74</td>
<td>0.81</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Sybil ranking

![Graph showing the fraction of Sybil nodes across different top K nodes]
Outline

1. Introduction to Sybil Attack
2. Background and Related Work
3. The SYBILFUSE Framework
4. Evaluation on Labeled Twitter Networks
5. Conclusion
Conclusion

- We proposed SYBILFUSE, a general framework that combines local attributes with global structure.
Conclusion

- We proposed SYBILFUSE, a general framework that combines local attributes with global structure.
- We evaluated SYBILFUSE on synthetic and real-world social networks, and demonstrated that SYBILFUSE outperforms existing approaches.
Thank you!

Q&A