Collective Spammer Detection in Evolving Multi-Relational Social Networks

Shobeir Fakhraei (University of Maryland)
James Foulds (University of California, Santa Cruz)
Madhusudana Shashanka (if(we) Inc., Currently Niara Inc.)
Lise Getoor (University of California, Santa Cruz)
Spam in Social Networks

- Recent study by Nexgate in 2013:
  - Spam grew by more than 300% in half a year
Spam in Social Networks

- Recent study by Nexgate in 2013:
  - Spam grew by more than 300% in half a year
  - 1 in 200 social messages are spam
Spam in Social Networks

Recent study by Nexgate in 2013:

- Spam grew by more than 300% in half a year
- 1 in 200 social messages are spam
- 5% of all social apps are spammy
Spam in Social Networks

What’s different about social networks?

- Spammers have more ways to interact with users
Spam in Social Networks

What’s different about social networks?

- Spammers have more ways to interact with users
- Messages, comments on photos, winks, …
Spam in Social Networks

- What’s different about social networks?
  - Spammers have more ways to interact with users
    - Messages, comments on photos, winks,…
  - They can split spam across multiple messages
What’s different about social networks?

- Spammers have more ways to interact with users
  - Messages, comments on photos, winks, …
- They can split spam across multiple messages
- More available info about users on their profiles!
Spammers are getting smarter!

**Traditional Spam:**

Want some replica luxury watches? Click here: http://SpammyLink.com

George

Shobeir
Spammers are getting smarter!

**Traditional Spam:**

George

Want some replica luxury watches? Click here: http://SpammyLink.com

Shobeir

[Report Spam]
Spammers are getting smarter!

Traditional Spam:
George

Want some replica luxury watches? Click here: http://SpammyLink.com

[Report Spam]

Shobeir

(Intelligent) Social Spam:
Mary
Hey Shobeir! Nice profile photo. I live in Bay Area too. Wanna chat?

Shobeir

Shobeir
Spammers are getting smarter!

Traditional Spam:

George

Want some replica luxury watches? Click here: http://SpammyLink.com

[Report Spam]

Shobeir

(Intelligent) Social Spam:

Mary

Hey Shobeir! Nice profile photo. I live in Bay Area too. Wanna chat?

Sure! :)

Shobeir
Spammers are getting smarter!

**Traditional Spam:**

George

Want some replica luxury watches? Click here: [http://SpammyLink.com](http://SpammyLink.com)

Shobeir

[Report Spam]

**(Intelligent) Social Spam:**

Mary

Hey Shobeir! Nice profile photo. I live in Bay Area too. Wanna chat?

Shobeir

Sure! :)

Mary

I’m logging off here., too many people pinging me! I really like you, let’s chat more here: [http://SpammyLink.com](http://SpammyLink.com)

Shobeir

Realistic Looking Conversation
Tagged.com

- Founded in 2004, is a social networking site which connects people through social interactions and games

- Over 300 million registered members

- Data sample for experiments (on a laptop):
  - 5.6 Million users (3.9% Labeled Spammers)
  - 912 Million Links
Social Networks: Multi-relational and Time-Evolving
Social Networks: Multi-relational and Time-Evolving

Legitimate users
Social Networks: Multi-relational and Time-Evolving

Legitimate users

Spammers
Social Networks: Multi-relational and Time-Evolving

**Link** = Action at time t

**Actions** = Profile view, message, poke, report abuse, etc
Social Networks: Multi-relational and Time-Evolving

Link = Action at time t

Actions = Profile view, message, poke, report abuse, etc
Social Networks: Multi-relational and Time-Evolving

**Link** = Action at time t

**Actions** = Profile view, message, poke, report abuse, etc
Social Networks: Multi-relational and Time-Evolving

**Link** = Action at time $t$

**Actions** = Profile view, message, poke, report abuse, etc
Social Networks: Multi-relational and Time-Evolving

**Link** = Action at time t

**Actions** = Profile view, message, poke, report abuse, etc
Social Networks: Multi-relational and Time-Evolving

**Link** = Action at time t

**Actions** = Profile view, message, poke, report abuse, etc
Our Approach

Predict spammers based on:

- Graph structure
- Action sequences
- Reporting behavior
Our Approach

Predict spammers based on:

- **Graph structure**
- Action sequences
- Reporting behavior
Graph Structure Feature Extraction

- Pagerank
- K-core
- Graph coloring
- Triangle count
- Connected components
- In/out degree

Graphs for each relation
Graph Structure Feature Extraction

Features

Graphs for each relation

- Pagerank, K-core, Graph coloring, Triangle count, Connected components, In/out degree
Graph Structure Features

- Extract features for each relation graph
  - PageRank
  - Degree statistics
    - Total degree
    - In degree
    - Out degree
  - k-Core
  - Graph coloring
  - Connected components
  - Triangle count

(8 features for each of 10 relations)
Graph Structure Features

- Extract features for each relation graph
  - **PageRank**
  - Degree statistics
    - Total degree
    - In degree
    - Out degree
  - k-Core
  - Graph coloring
  - Connected components
  - Triangle count

(8 features for each of 10 relations)
Graph Structure Features

- Extract features for each relation graph
  - PageRank
  - **Degree statistics**
    - Total degree
    - In degree
    - Out degree
  - k-Core
  - Graph coloring
  - Connected components
  - Triangle count

(8 features for each of 10 relations)
Graph Structure Features

- Extract features for each relation graph
  - PageRank
  - Degree statistics
    - Total degree
    - In degree
    - Out degree
  - k-Core
  - Graph coloring
  - Connected components
  - Triangle count

(8 features for each of 10 relations)
Graph Structure Features

- Extract features for each relation graph
  - PageRank
  - Degree statistics
    - Total degree
    - In degree
    - Out degree
  - k-Core
  - Graph coloring
  - Connected components
  - Triangle count

(8 features for each of 10 relations)
Graph Structure Features

- Extract features for each relation graph
  - PageRank
  - Degree statistics
    - Total degree
    - In degree
    - Out degree
  - k-Core
  - Graph coloring
  - Connected components
  - Triangle count

(8 features for each of 10 relations)
Graph Structure Features

- Extract features for each relation graph
  - PageRank
  - Degree statistics
    - Total degree
    - In degree
    - Out degree
  - k-Core
  - Graph coloring
  - Connected components
  - Triangle count

(8 features for each of 10 relations)
Graph Structure Features

- Extract features for each relation graph
  - PageRank
  - Degree statistics
    - Total degree
    - In degree
    - Out degree
  - k-Core
  - Graph coloring
  - Connected components
  - Triangle count

(8 features for each of 10 relations)
Graph Structure Features

- Extract features for each relation graph

  - PageRank
  - Degree statistics
    - Total degree
    - In degree
    - Out degree
  - k-Core
  - Graph coloring
  - Connected components
  - Triangle count

(8 features for each of 10 relations)
Graph Structure Features

Classification method: Gradient Boosted Trees
### Graph Structure Features

<table>
<thead>
<tr>
<th>Experiments</th>
<th>AU-PR</th>
<th>AU-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Relation, 8 Feature types</td>
<td>0.187 ± 0.004</td>
<td>0.803 ± 0.001</td>
</tr>
<tr>
<td>10 Relations, 1 Feature type</td>
<td>0.285 ± 0.002</td>
<td>0.809 ± 0.001</td>
</tr>
<tr>
<td>10 Relations, 8 Feature types</td>
<td>0.328 ± 0.003</td>
<td>0.817 ± 0.001</td>
</tr>
</tbody>
</table>

Multiple relations/features → better performance!
## Graph Structure Features

<table>
<thead>
<tr>
<th>Experiments</th>
<th>AU-PR</th>
<th>AU-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Relation, 8 Feature types</td>
<td>0.187 ± 0.004</td>
<td>0.803 ± 0.001</td>
</tr>
<tr>
<td>10 Relations, 1 Feature type</td>
<td>0.285 ± 0.002</td>
<td>0.809 ± 0.001</td>
</tr>
<tr>
<td>10 Relations, 8 Feature types</td>
<td>0.328 ± 0.003</td>
<td>0.817 ± 0.001</td>
</tr>
</tbody>
</table>

Multiple relations/features lead to better performance!
## Graph Structure Features

<table>
<thead>
<tr>
<th>Experiments</th>
<th>AU-PR</th>
<th>AU-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Relation, 8 Feature types</td>
<td>0.187 ± 0.004</td>
<td>0.803 ± 0.001</td>
</tr>
<tr>
<td>10 Relations, 1 Feature type</td>
<td>0.285 ± 0.002</td>
<td>0.809 ± 0.001</td>
</tr>
<tr>
<td>10 Relations, 8 Feature types</td>
<td>0.328 ± 0.003</td>
<td>0.817 ± 0.001</td>
</tr>
</tbody>
</table>

Multiple relations/features → better performance!
## Graph Structure Features

<table>
<thead>
<tr>
<th>Experiments</th>
<th>AU-PR</th>
<th>AU-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Relation, 8 Feature types</td>
<td>0.187 ± 0.004</td>
<td>0.803 ± 0.001</td>
</tr>
<tr>
<td>10 Relations, 1 Feature type</td>
<td>0.285 ± 0.002</td>
<td>0.809 ± 0.001</td>
</tr>
<tr>
<td>10 Relations, 8 Feature types</td>
<td>0.328 ± 0.003</td>
<td>0.817 ± 0.001</td>
</tr>
</tbody>
</table>

Multiple relations/features → better performance!
Our Approach

Predict spammers based on:

- Graph structure
- **Action sequences**
- Reporting behavior
Sequence of Actions

- **Sequential Bigram Features:**
  Short sequence segment of 2 consecutive actions, to capture sequential information

User1 Actions:
Message, Profile_view, Message, Friend_Request, ....
Sequence of Actions

- **Mixture of Markov Models (MMM):**
  A.k.a. chain-augmented, tree-augmented naive Bayes

\[ P(y, x) = P(y)P(x_1 | y) \prod_{i=2}^{n} P(x_i | x_{i-1}, y), \]
Sequence of Actions

Action Sequence

Bigram Features

Chain Augmented NB

$P(y, x) = P(y)P(x_1 | y) \prod_{i=2}^{n} P(x_i | x_{i-1}, y)$,
### Sequence of Actions

<table>
<thead>
<tr>
<th>Experiments</th>
<th>AU-PR</th>
<th>AU-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bigram Features</td>
<td>0.471 ± 0.004</td>
<td>0.859 ± 0.001</td>
</tr>
<tr>
<td>MMM</td>
<td>0.246 ± 0.009</td>
<td>0.821 ± 0.003</td>
</tr>
<tr>
<td>Bigram + MMM</td>
<td>0.468 ± 0.012</td>
<td>0.860 ± 0.002</td>
</tr>
</tbody>
</table>

Little benefit from MMM (although little overhead)
We can classify 70% of the spammers that need manual labeling with about 90% accuracy
Deployment and Example Runtimes

- **We can:**
  - Run the model on short intervals, with new snapshots of the network
  - Update the features as events occur

- **Example runtimes with Graphlab Create™ on a Macbook Pro:**
  - 5.6 million vertices and 350 million edges:
    - PageRank: 6.25 minutes
    - Triangle counting: 17.98 minutes
    - k-core: 14.3 minutes
Our Approach

Predict spammers based on:

- Graph structure
- Action sequences
- **Reporting behavior**
Refining the abuse reporting systems

- Abuse report systems are very noisy
  - People have different standards
  - Spammers report random people to increase noise
  - Personal gain in social games

- Goal is to clean up the system using:
  - Reporters’ previous history
  - Collective reasoning over reports
Collective Classification with Reports

Probabilistic Soft Logic

- \( \text{CREDIBLE}(v_1) \land \text{REPORTED}(v_1, v_2) \rightarrow \text{SPAMMER}(v_2) \)
- \( \text{SPAMMER}(v_2) \land \text{REPORTED}(v_1, v_2) \rightarrow \text{CREDIBLE}(v_1) \)
- \( \neg\text{SPAMMER}(v_2) \land \text{REPORTED}(v_1, v_2) \rightarrow \neg\text{CREDIBLE}(v_1) \)
- \( \text{PRIOR-CREDIBLE}(v) \rightarrow \text{CREDIBLE}(v) \)
- \( \neg\text{PRIOR-CREDIBLE}(v) \rightarrow \neg\text{CREDIBLE}(v) \)
- \( \neg\text{SPAMMER}(v) \)
HL-MRFs & Probabilistic Soft Logic (PSL)

- Probabilistic Soft Logic (PSL), a declarative modeling language based on first-order logic

- Weighted logical rules define a probabilistic graphical model:

  \[ \omega : P(A, B) \land Q(B, C') \rightarrow R(A, C') \]

- Instantiated rules reduce the probability of any state that does not satisfy the rule, as measured by its distance to satisfaction
Collective Classification with Reports

- Model using only reports:

\[ \text{REPORTED}(v_1, v_2) \rightarrow \text{SPAMMER}(v_2) \]
\[ \neg \text{SPAMMER}(v) \]
Collective Classification with Reports

- Model using reports and credibility of the reporter:

\[ CREDIBLE(v_1) \land REPORTED(v_1, v_2) \rightarrow SPAMMER(v_2) \]
\[ PRIOR-CREDIBLE(v) \rightarrow CREDIBLE(v) \]
\[ \neg PRIOR-CREDIBLE(v) \rightarrow \neg CREDIBLE(v) \]
\[ \neg SPAMMER(v) \]
Collective Classification with Reports

- Model using reports, credibility of the reporter, and collective reasoning:

\[
\begin{align*}
CREDIBLE(v_1) \land REPORTED(v_1, v_2) & \rightarrow SPAMMER(v_2) \\
SPAMMER(v_2) \land REPORTED(v_1, v_2) & \rightarrow CREDIBLE(v_1) \\
\neg SPAMMER(v_2) \land REPORTED(v_1, v_2) & \rightarrow \neg CREDIBLE(v_1) \\
PRIOR-CREDIBLE(v) & \rightarrow CREDIBLE(v) \\
\neg PRIOR-CREDIBLE(v) & \rightarrow \neg CREDIBLE(v) \\
\neg SPAMMER(v) & \rightarrow \neg CREDIBLE(v)
\end{align*}
\]
Results of Classification Using Reports

<table>
<thead>
<tr>
<th>Experiments</th>
<th>AU-PR</th>
<th>AU-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reports Only</td>
<td>0.674 ± 0.008</td>
<td>0.611 ± 0.007</td>
</tr>
<tr>
<td>Reports &amp; Credibility</td>
<td>0.869 ± 0.006</td>
<td>0.862 ± 0.004</td>
</tr>
<tr>
<td>Reports &amp; Credibility &amp; Collective Reasoning</td>
<td>0.884 ± 0.005</td>
<td>0.873 ± 0.004</td>
</tr>
</tbody>
</table>
### Results of Classification Using Reports

<table>
<thead>
<tr>
<th>Experiments</th>
<th>AU-PR</th>
<th>AU-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reports Only</td>
<td>0.674 ± 0.008</td>
<td>0.611 ± 0.007</td>
</tr>
<tr>
<td>Reports &amp; Credibility</td>
<td>0.869 ± 0.006</td>
<td>0.862 ± 0.004</td>
</tr>
<tr>
<td>Reports &amp; Credibility &amp; Collective Reasoning</td>
<td>0.884 ± 0.005</td>
<td>0.873 ± 0.004</td>
</tr>
</tbody>
</table>
# Results of Classification Using Reports

<table>
<thead>
<tr>
<th>Experiments</th>
<th>AU-PR</th>
<th>AU-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reports Only</td>
<td>0.674 ± 0.008</td>
<td>0.611 ± 0.007</td>
</tr>
<tr>
<td>Reports &amp; Credibility</td>
<td>0.869 ± 0.006</td>
<td>0.862 ± 0.004</td>
</tr>
<tr>
<td>Reports &amp; Credibility &amp; Collective Reasoning</td>
<td>0.884 ± 0.005</td>
<td>0.873 ± 0.004</td>
</tr>
</tbody>
</table>
# Results of Classification Using Reports

<table>
<thead>
<tr>
<th>Experiments</th>
<th>AU-PR</th>
<th>AU-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reports Only</td>
<td>0.674 ± 0.008</td>
<td>0.611 ± 0.007</td>
</tr>
<tr>
<td>Reports &amp; Credibility</td>
<td>0.869 ± 0.006</td>
<td>0.862 ± 0.004</td>
</tr>
<tr>
<td>Reports &amp; Credibility &amp; Collective Reasoning</td>
<td>0.884 ± 0.005</td>
<td>0.873 ± 0.004</td>
</tr>
</tbody>
</table>
Conclusion

Multiple relations are more predictive than multiple features

AUPR: 0.187 → 0.328

Code and part of the data will be released soon:
https://github.com/shobeir/fakhraei_kdd2015
Conclusion

Multiple relations are more predictive than multiple features

AUPR: 
0.187 → 0.328

Even simple bigrams are highly predictive

AUPR: 0.471

Code and part of the data will be released soon:
https://github.com/shobeir/fakhraei_kdd2015
Conclusion

Multiple relations are more predictive than multiple features
AUPR: 0.187 → 0.328

Can classify 70% of the spammers that needed manual labeling with 90% accuracy
AUPR: 0.779

Even simple bigrams are highly predictive
AUPR: 0.471

Even simple bigrams are highly predictive

Code and part of the data will be released soon:
https://github.com/shobeir/fakhraei_kdd2015
Conclusion

Multiple relations are more predictive than multiple features

AUPR: $0.187 \rightarrow 0.328$

Can classify 70% of the spammers that needed manual labeling with 90% accuracy

AUPR: 0.779

Even simple bigrams are highly predictive

AUPR: 0.471

Jointly refining the credibility of the source is highly effective!

AUPR: $0.674 \rightarrow 0.884$

AUPR: $0.471$

Can classify 70% of the spammers that needed manual labeling with 90% accuracy

AUPR: 0.779

Even simple bigrams are highly predictive

AUPR: 0.471

Jointly refining the credibility of the source is highly effective!

AUPR: $0.674 \rightarrow 0.884$

Code and part of the data will be released soon:
https://github.com/shobeir/fakhraei_kdd2015
Acknowledgements

**Collaborators:**

- Shobeir Fakhraei  
  *Univ. of Maryland*
- Madhusudana Shashanka  
  *if(we) Inc., currently Niara Inc.*
- Lise Getoor  
  *Univ. California, Santa Cruz*

**If(we) Inc. (Formerly Tagged Inc.):**

- Johann Schleier-Smith
- Karl Dawson
- Dai Li
- Stuart Robinson
- Vinit Garg
- Simon Hill

**Dato (Formerly Graphlab):**

- Danny Bickson
- Brian Kent
- Srikrishna Sridhar
- Rajat Arya
- Shawn Scully
- Alice Zheng
Conclusion

Multiple relations are more predictive than multiple features

Can classify 70% of the spammers that needed manual labeling with 90% accuracy

Even simple bigrams are highly predictive

AUPR: 0.779

Jointly refining the credibility of the source is highly effective!

AUPR: 0.674 → 0.884

Can classify 70% of the spammers that needed manual labeling with 90% accuracy

AUPR: 0.187 → 0.328

AUPR: 0.471

AUPR: 0.674 → 0.884

Can classify 70% of the spammers that needed manual labeling with 90% accuracy

AUPR: 0.187 → 0.328

AUPR: 0.471

AUPR: 0.674 → 0.884

Code and part of the data will be released soon:
https://github.com/shobeir/fakhraei_kdd2015