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

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
  - More available info about users on their profiles!

Spammers are getting smarter!

Traditional Spam:

George

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

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

Shobeir

[Report Spam]

(Intelligent) Social Spam:

Mary

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

Shobeir

Spammers are getting smarter!

Traditional Spam:

George

Want some replica luxury watches?
Click here:
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! :)

Spammers are getting smarter!

Traditional Spam:

George

Want some replica luxury watches?
Click here:
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! :)

Realistic Looking Conversation ⋮

Mary

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

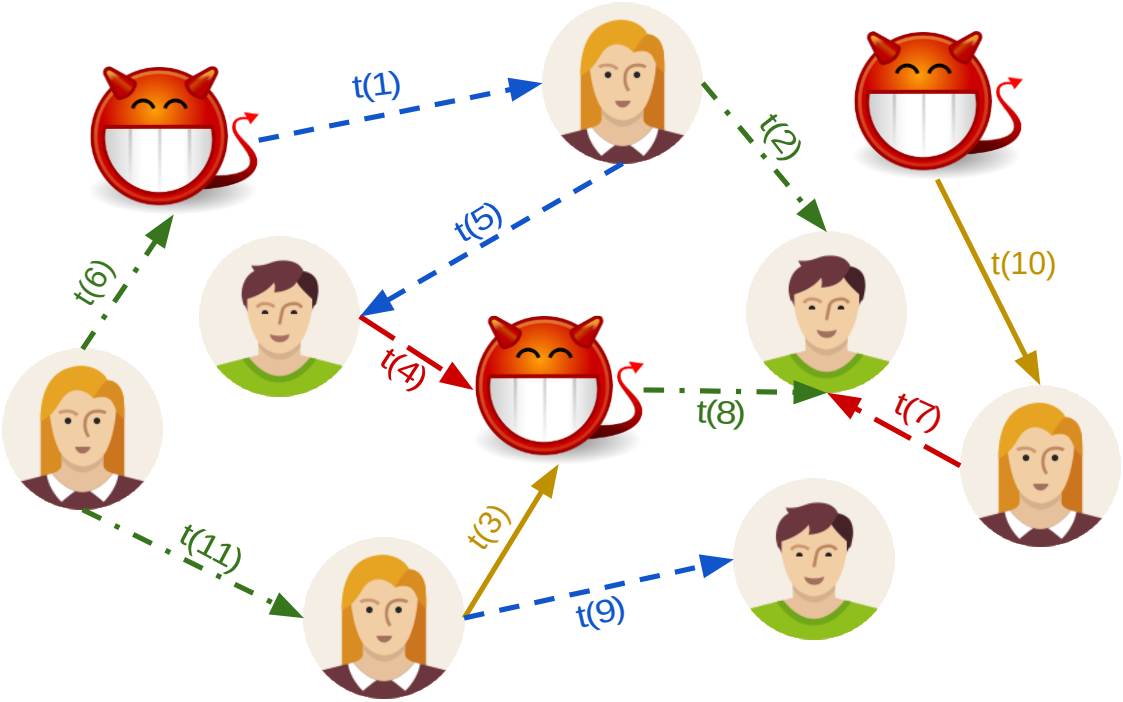
Tagged.com

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

- Over 300 million registered members

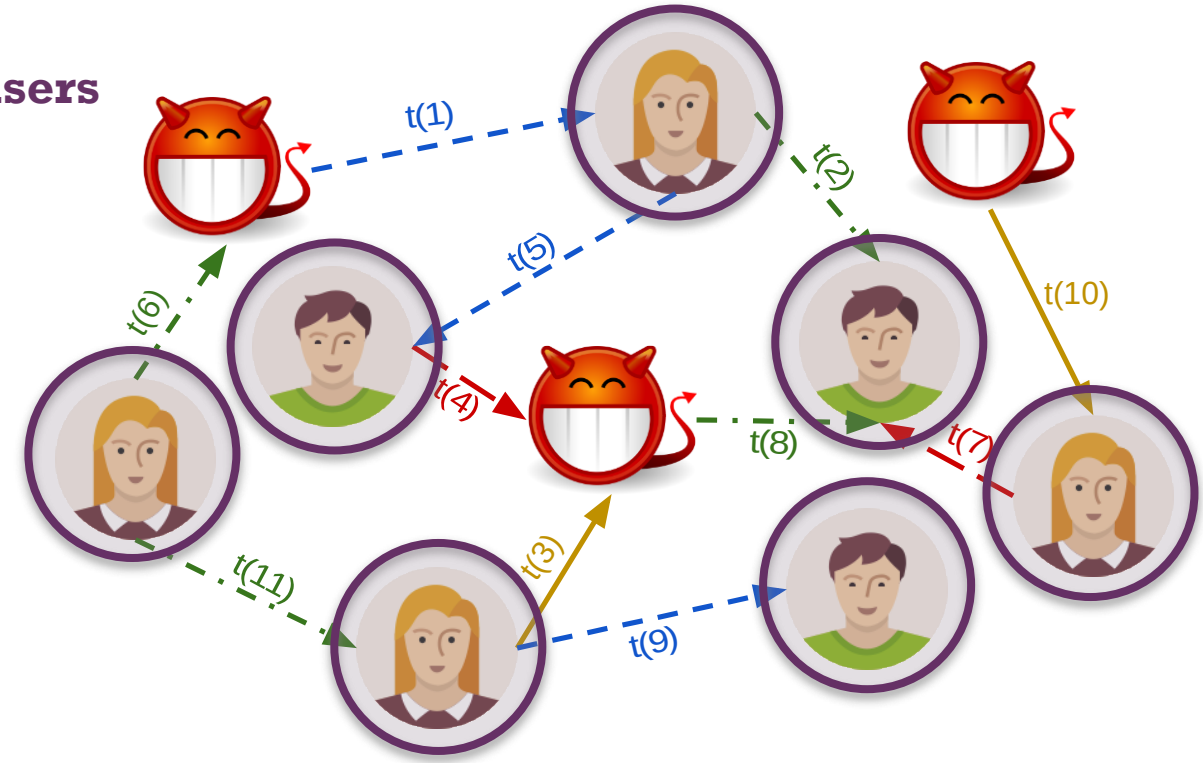
- Data sample for experiments (on a laptop):
  - 5.6 Million users (3.9% Labeled Spammers)
  - 912 Million Links

Social Networks:
Multi-relational and Time-Evolving



Social Networks: Multi-relational and Time-Evolving

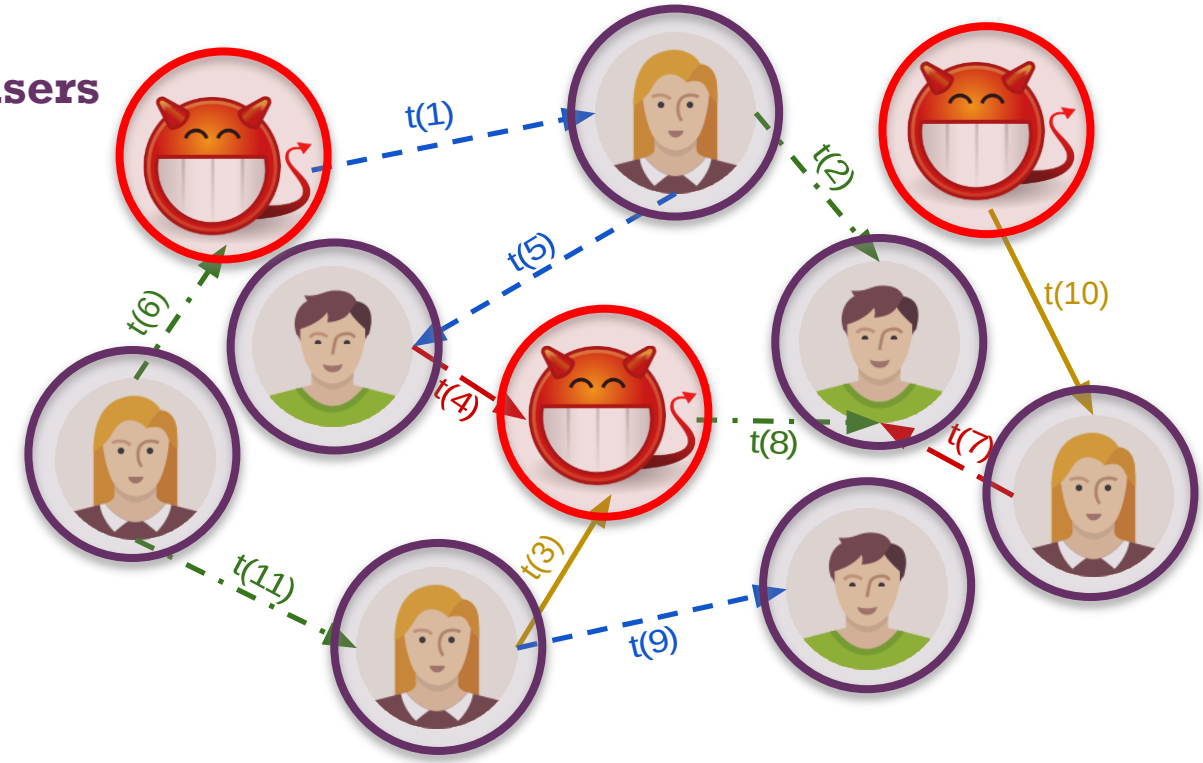
Legitimate users



Social Networks: Multi-relational and Time-Evolving

Legitimate users

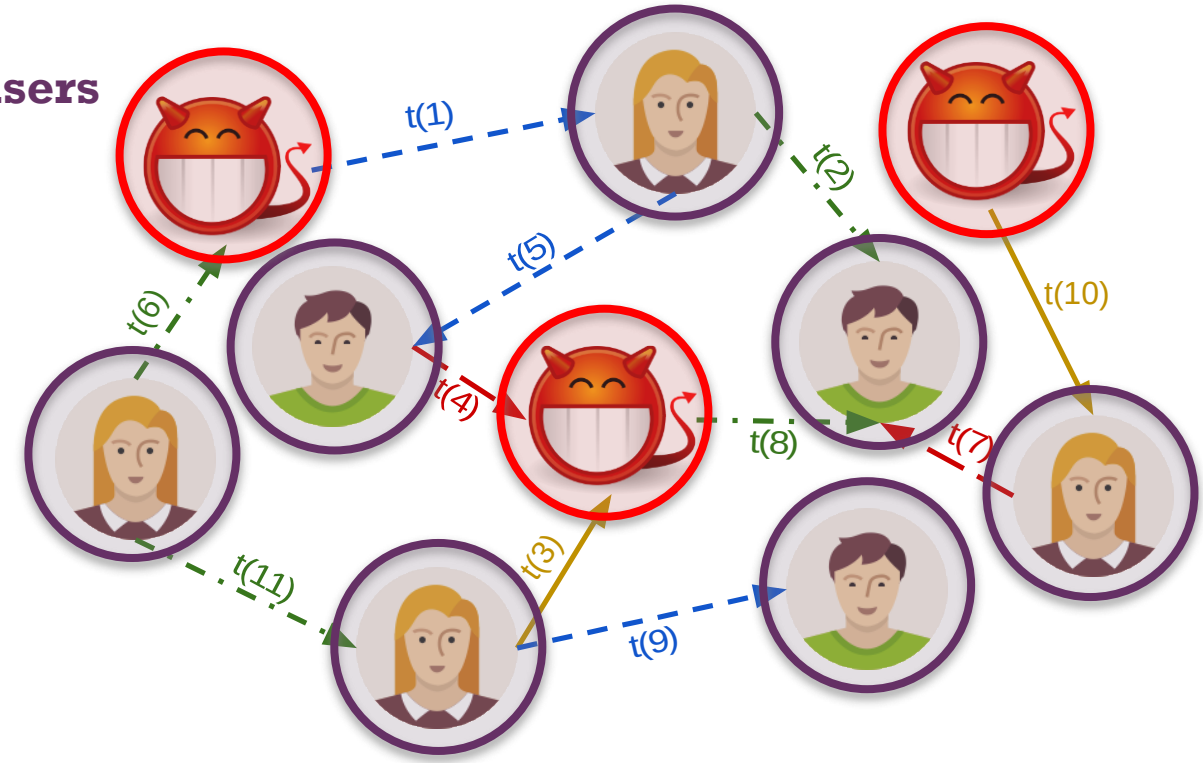
Spammers



Social Networks: Multi-relational and Time-Evolving

Legitimate users

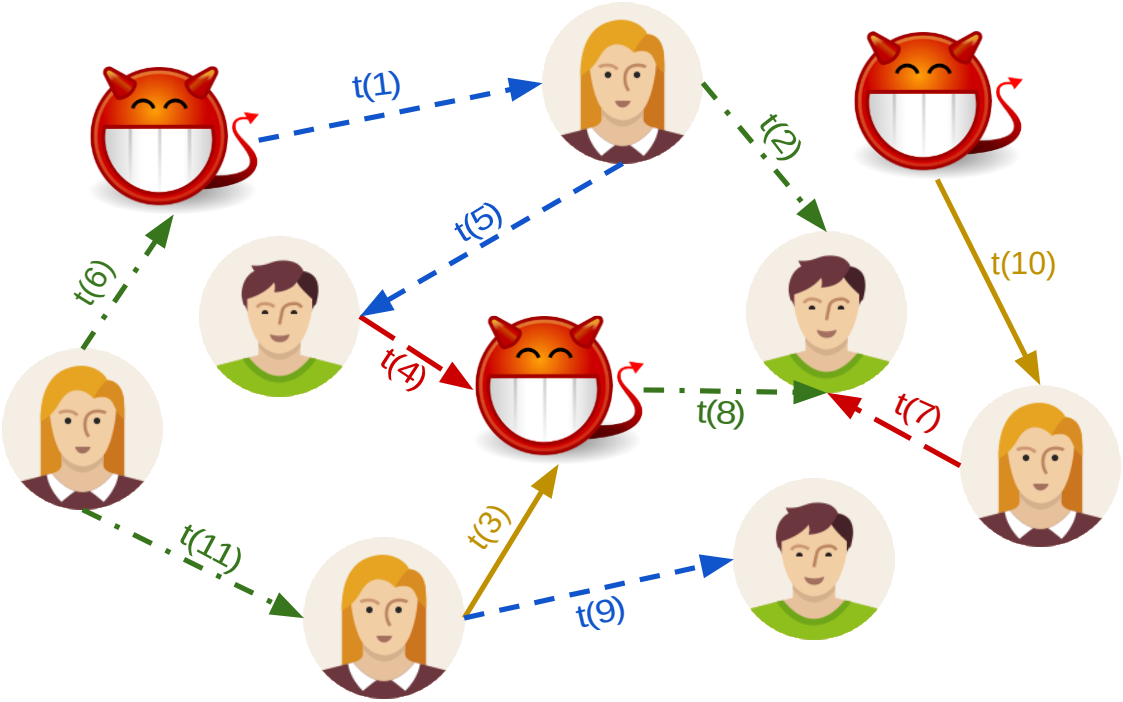
Spammers



Link = Action at time t

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

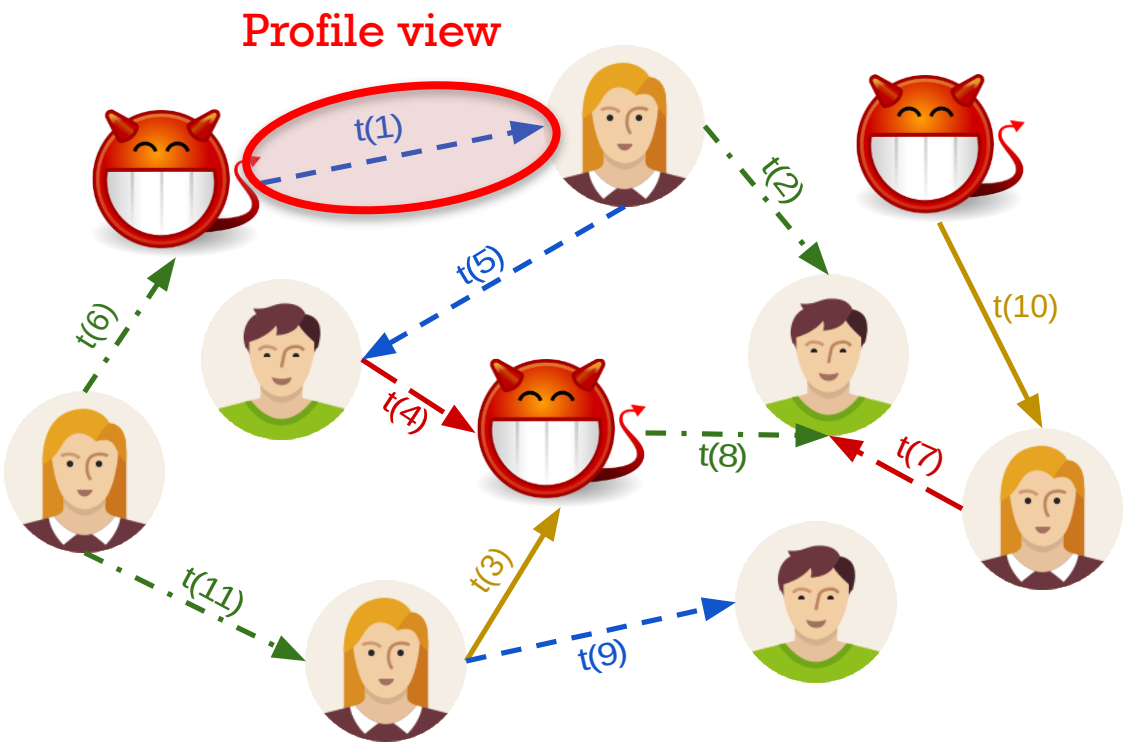
Social Networks: Multi-relational and Time-Evolving



Link = Action at time t

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

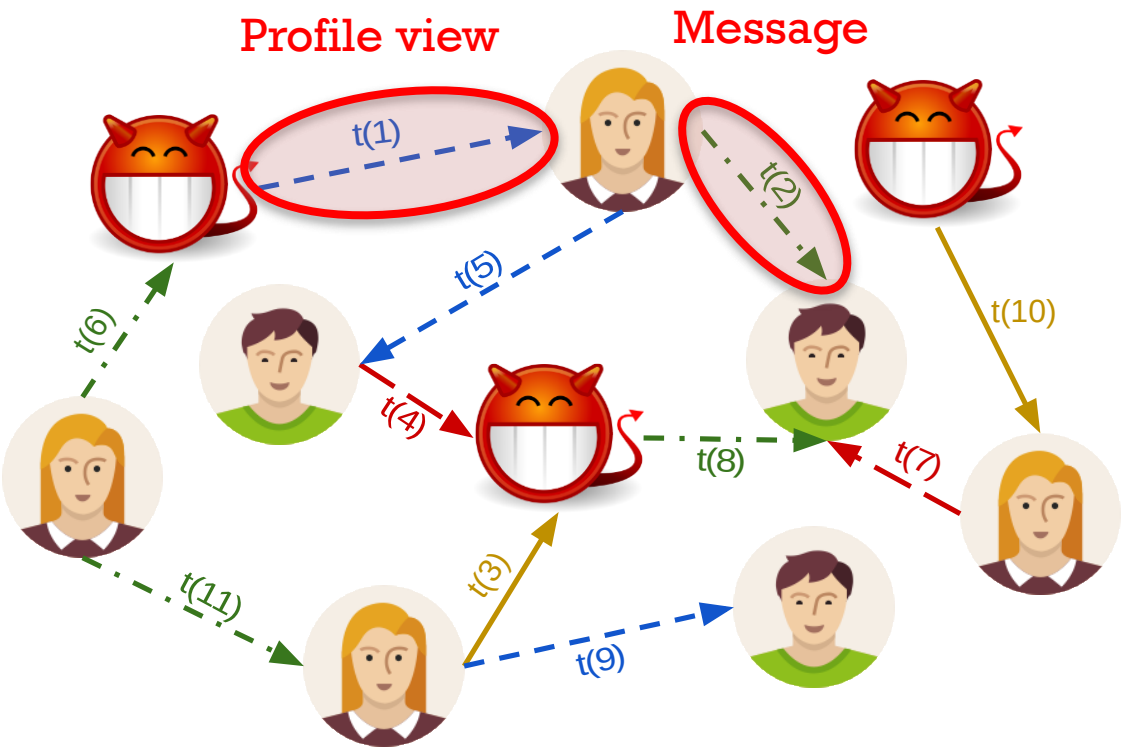
Social Networks: Multi-relational and Time-Evolving



Link = Action at time t

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

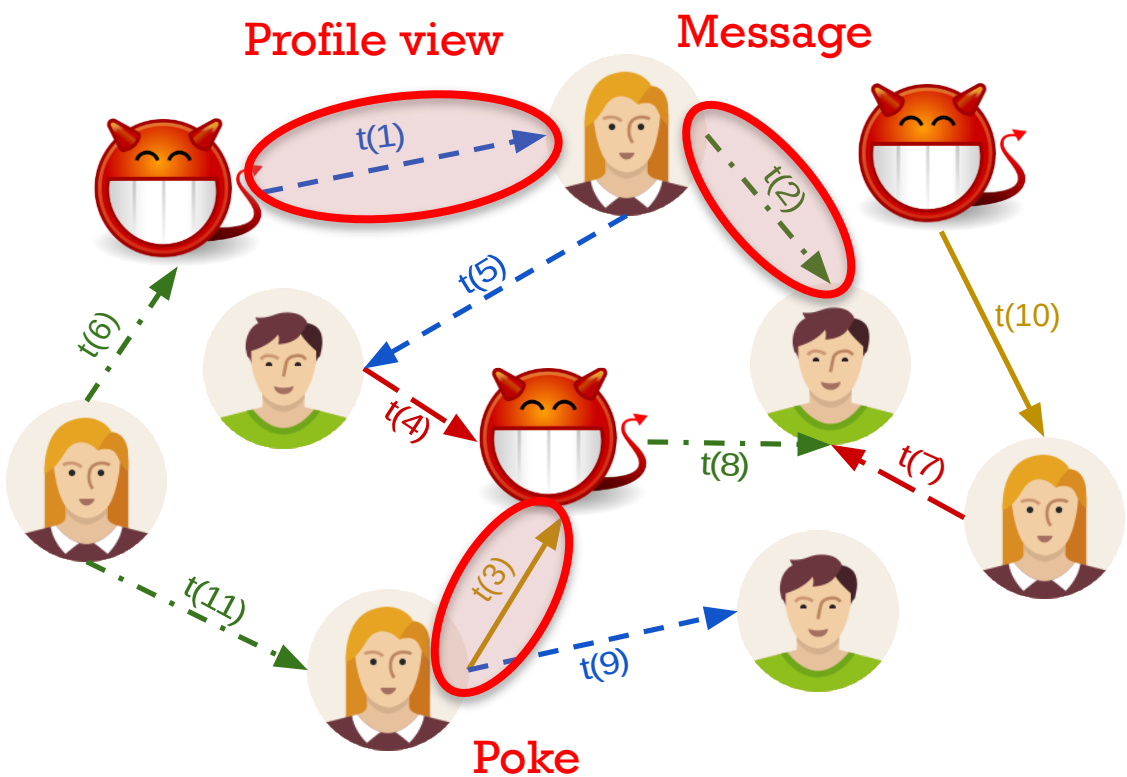
Social Networks: Multi-relational and Time-Evolving



Link = Action at time t

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

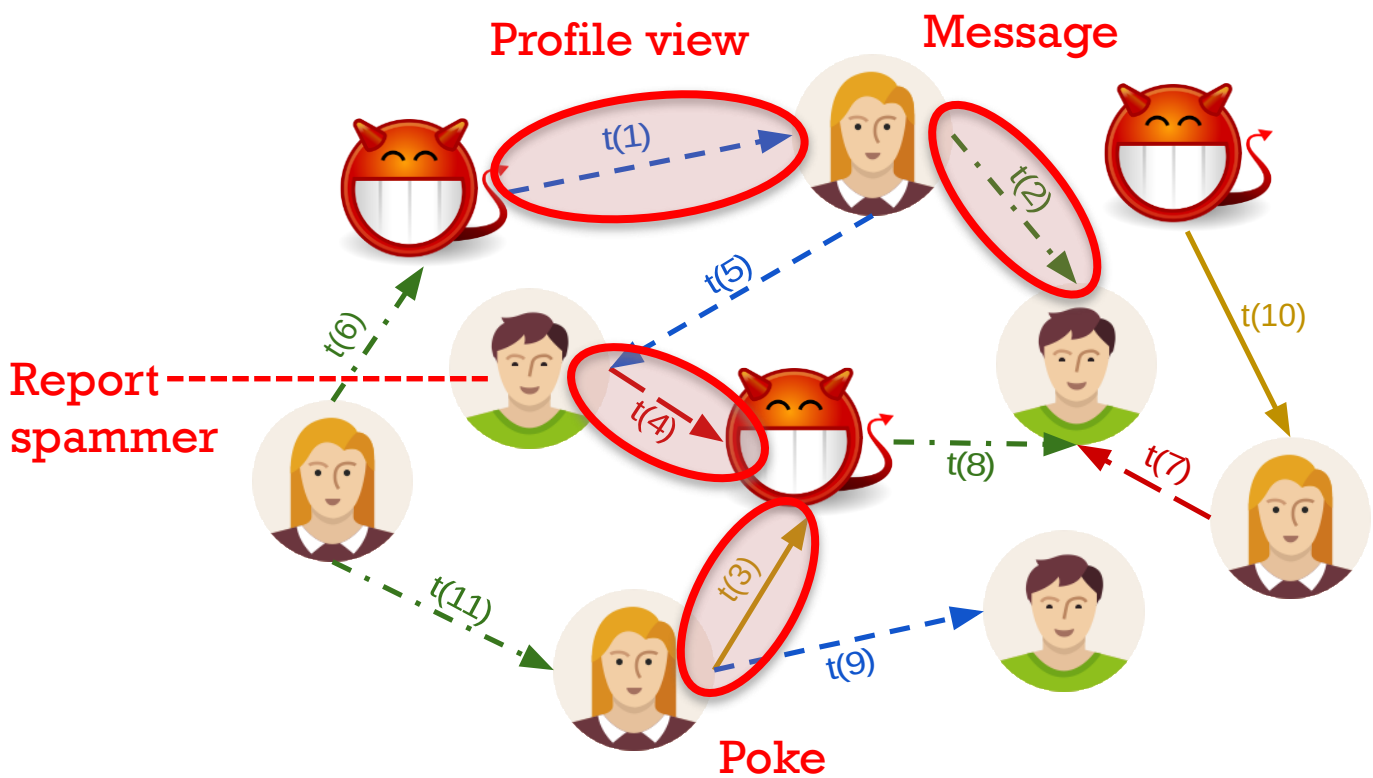
Social Networks: Multi-relational and Time-Evolving



Link = Action at time t

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

Social Networks: Multi-relational and Time-Evolving



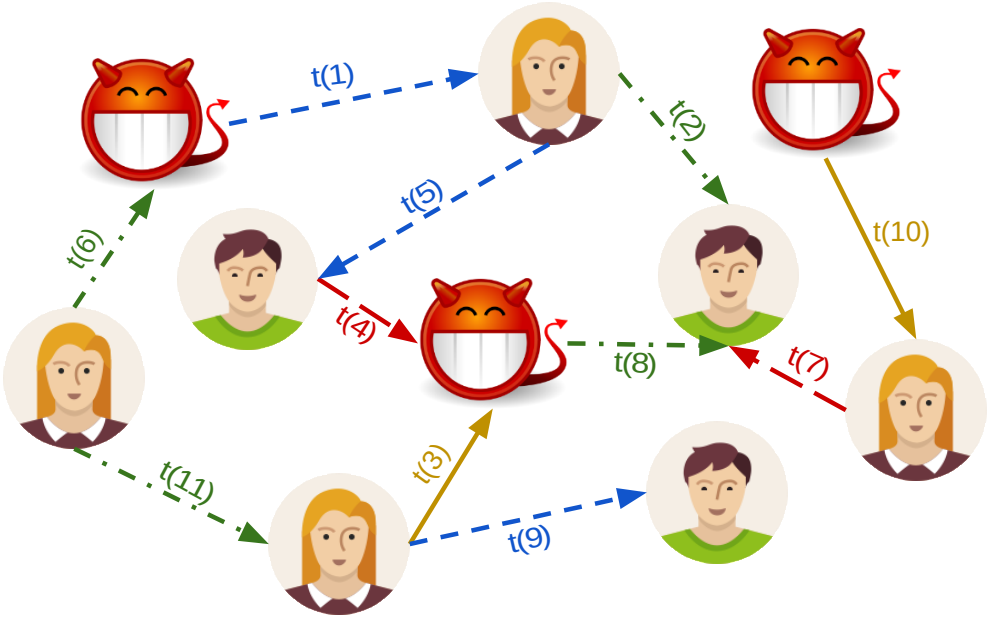
Link = Action at time t

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

Our Approach

Predict spammers based on:

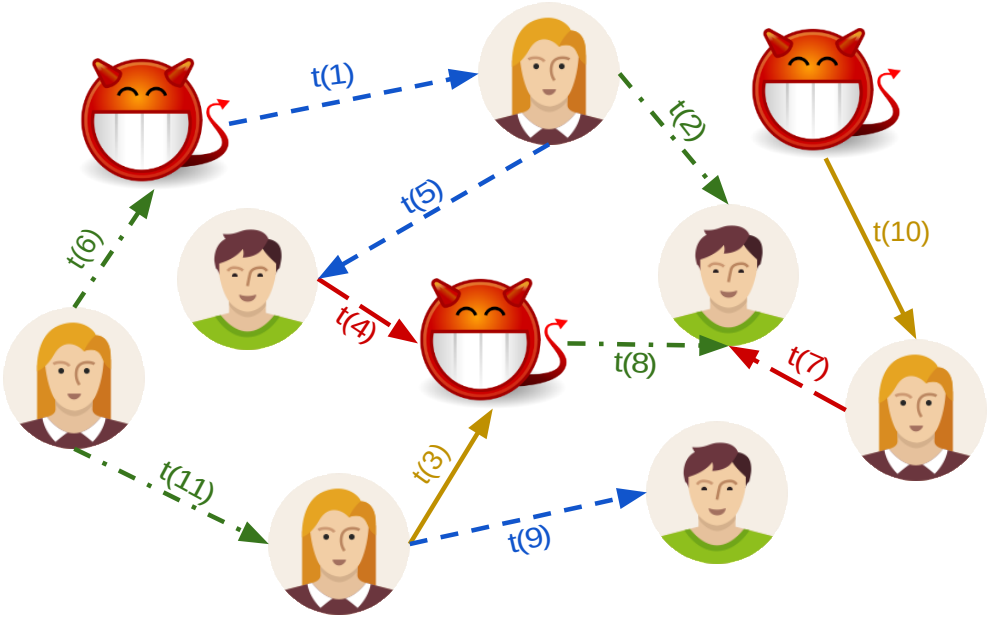
- Graph structure
- Action sequences
- Reporting behavior



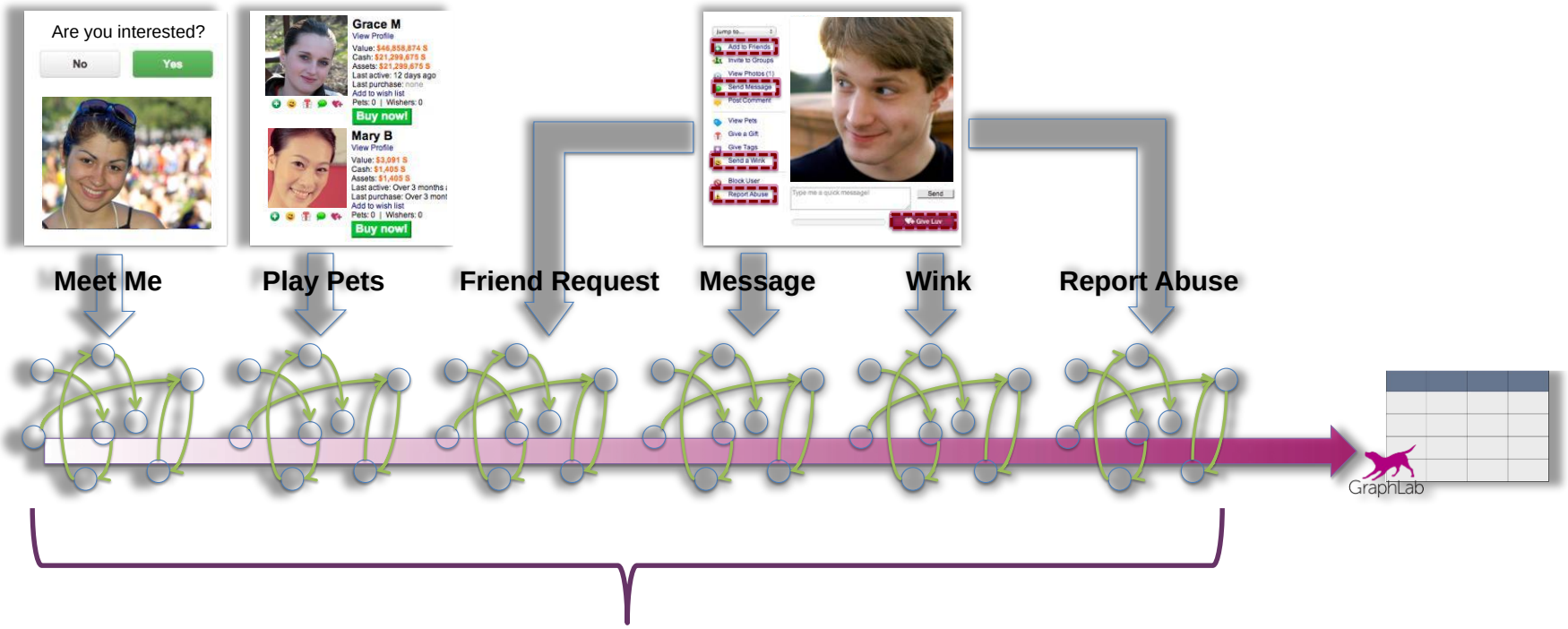
Our Approach

Predict spammers based on:

- **Graph structure**
- Action sequences
- Reporting behavior

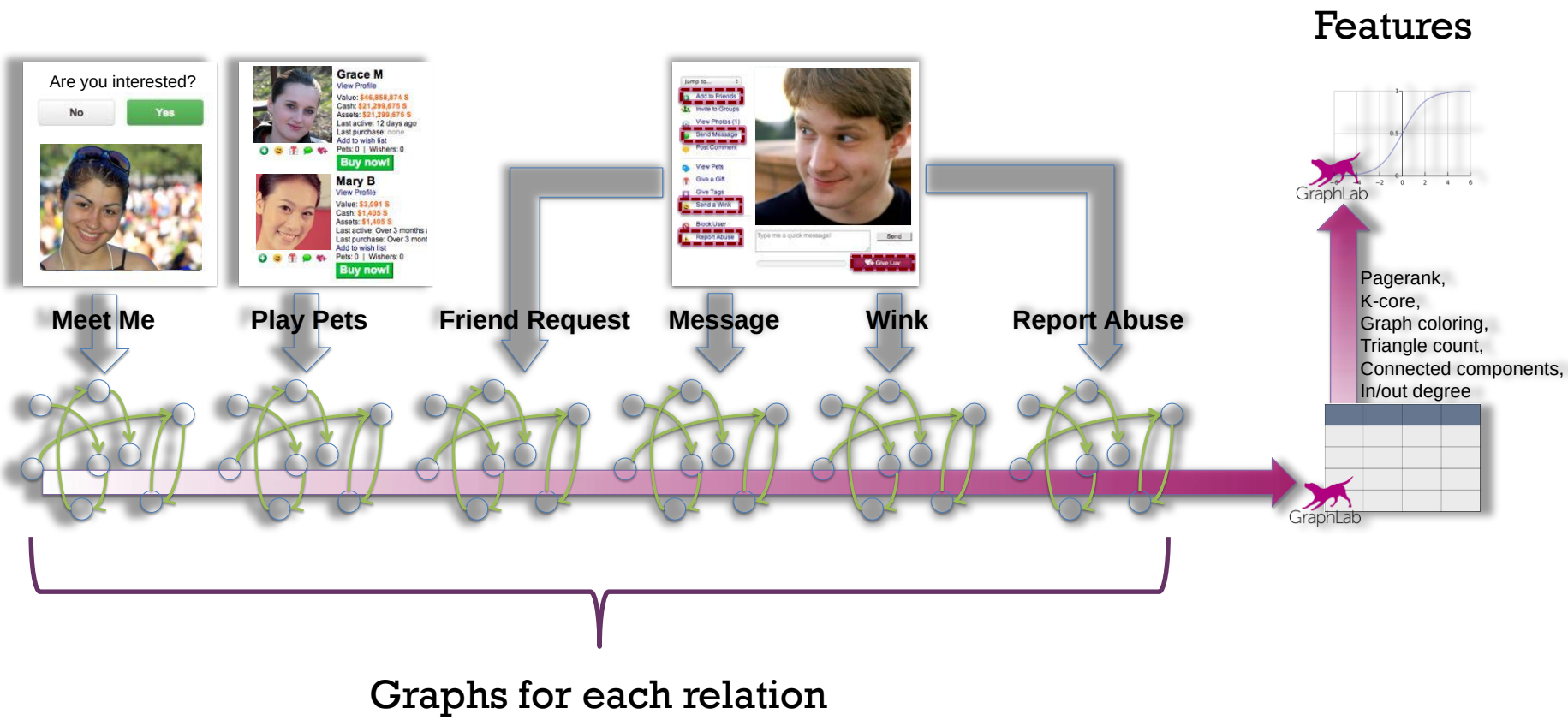


Graph Structure Feature Extraction



Graphs for each relation

Graph Structure Feature Extraction



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

X

(8 features for each of 10 relations)

Graph Structure Features

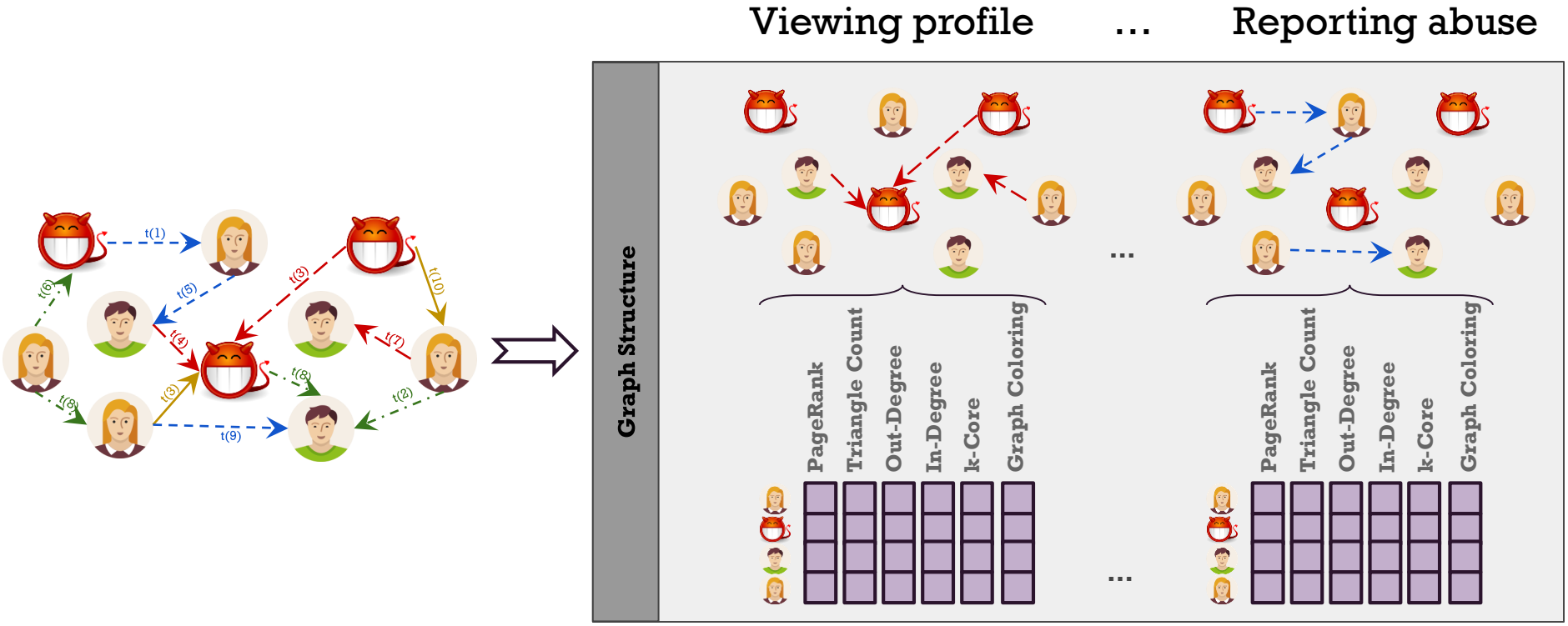
- Extract features for each relation graph
  - PageRank
  - Degree statistics
    - Total degree
    - In degree
    - Out degree
  - k-Core
  - Graph coloring
  - Connected components
  - Triangle count

X

- Viewing profile
- Friend requests
- Message
- Luv
- Wink
- Pets game
  - Buying
  - Wishing
- MeetMe game
  - Yes
  - No
- Reporting abuse

(8 features for each of 10 relations)

Graph Structure Features



Classification method: Gradient Boosted Trees

Graph Structure Features

Experiments	AU-PR	AU-ROC
1 Relation, 8 Feature types	0.187 ± 0.004	0.803 ±0.001
10 Relations, 1 Feature type	0.285 ± 0.002	0.809 ± 0.001
10 Relations, 8 Feature types	0.328 ± 0.003	0.817 ± 0.001

Multiple relations/features ⟹ better performance!

Graph Structure Features

Experiments	AU-PR	AU-ROC
1 Relation, 8 Feature types	**0.187 ± 0.004**	**0.803 ±0.001**
10 Relations, 1 Feature type	0.285 ± 0.002	0.809 ± 0.001
10 Relations, 8 Feature types	0.328 ± 0.003	0.817 ± 0.001

Multiple relations/features ⟹ better performance!

Graph Structure Features

Experiments	AU-PR	AU-ROC
1 Relation, 8 Feature types	0.187 ± 0.004	0.803 ±0.001
10 Relations, 1 Feature type	**0.285 ± 0.002**	**0.809 ± 0.001**
10 Relations, 8 Feature types	0.328 ± 0.003	0.817 ± 0.001

Multiple relations/features ⟹ better performance!

Graph Structure Features

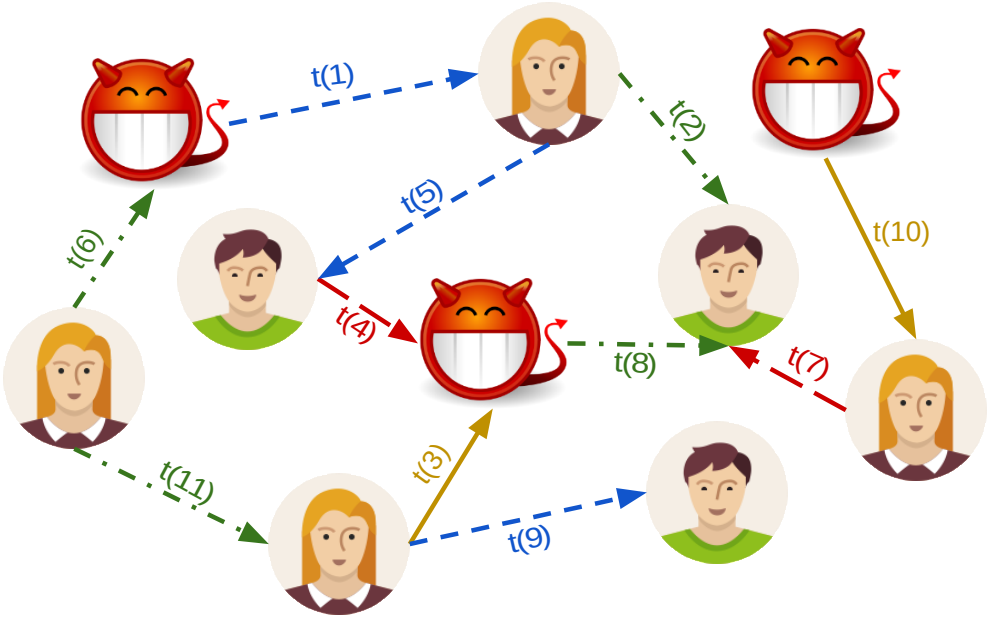
Experiments	AU-PR	AU-ROC
1 Relation, 8 Feature types	0.187 ± 0.004	0.803 ±0.001
10 Relations, 1 Feature type	0.285 ± 0.002	0.809 ± 0.001
10 Relations, 8 Feature types	**0.328 ± 0.003**	**0.817 ± 0.001**

Multiple relations/features ⟹ better performance!

Our Approach

Predict spammers based on:

- Graph structure
- **Action sequences**
- Reporting behavior



Sequence of Actions

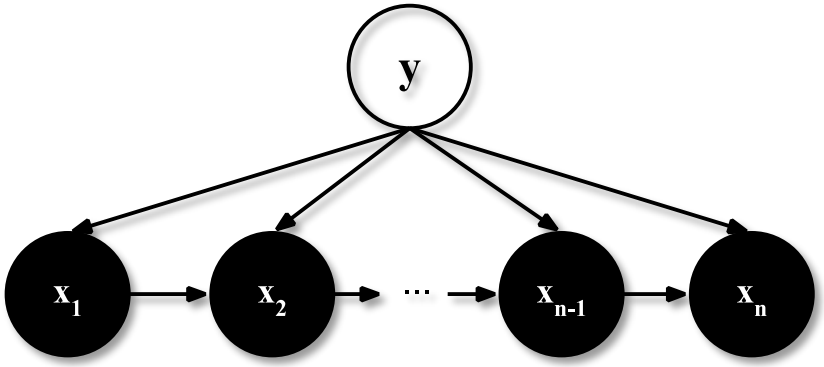
- **Sequential Bigram Features:**
  Short sequence segment of 2 consecutive actions, to capture sequential information

User1 Actions:
Message, Profile_view, Message, Friend_Request, ….

Sequence of Actions

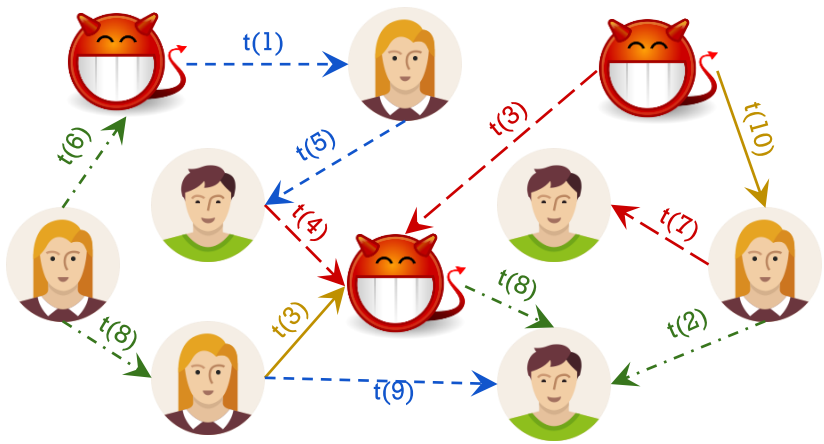
- **Mixture of Markov Models (MMM):**

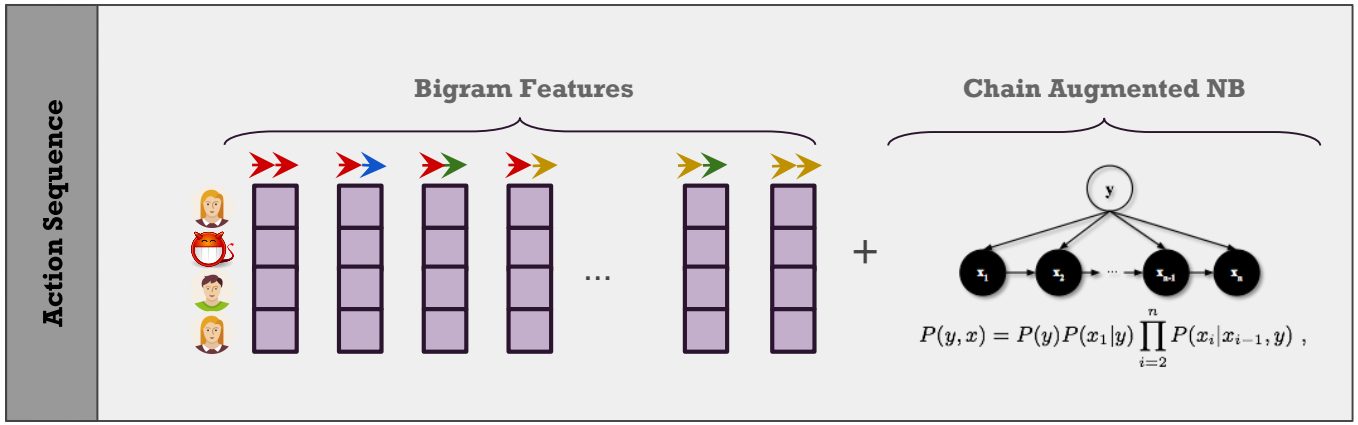
  A.k.a. chain-augmented, tree-augmented naive Bayes



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

Sequence of Actions





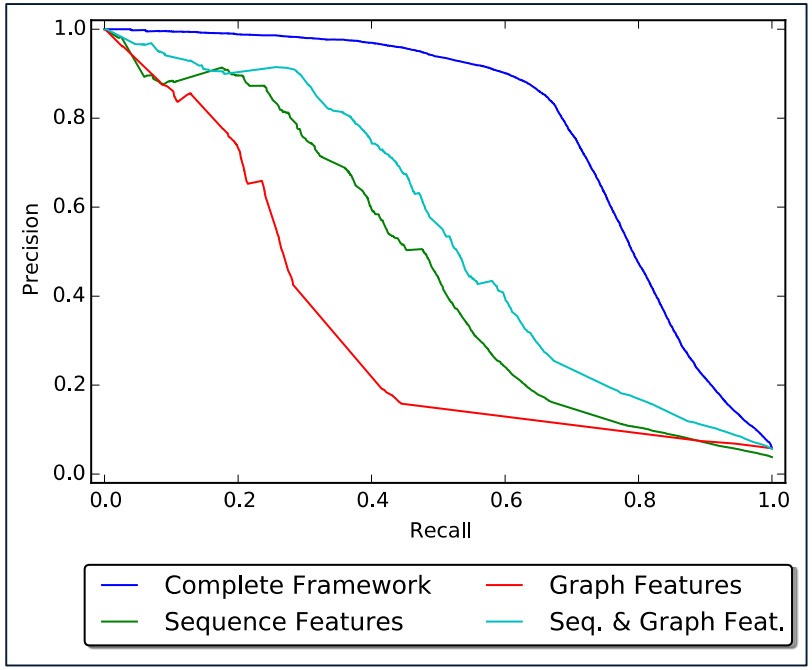
Sequence of Actions

Experiments	AU-PR	AU-ROC
Bigram Features	0.471 ± 0.004	0.859 ± 0.001
MMM	0.246 ± 0.009	0.821 ± 0.003
Bigram + MMM	0.468 ± 0.012	0.860 ± 0.002

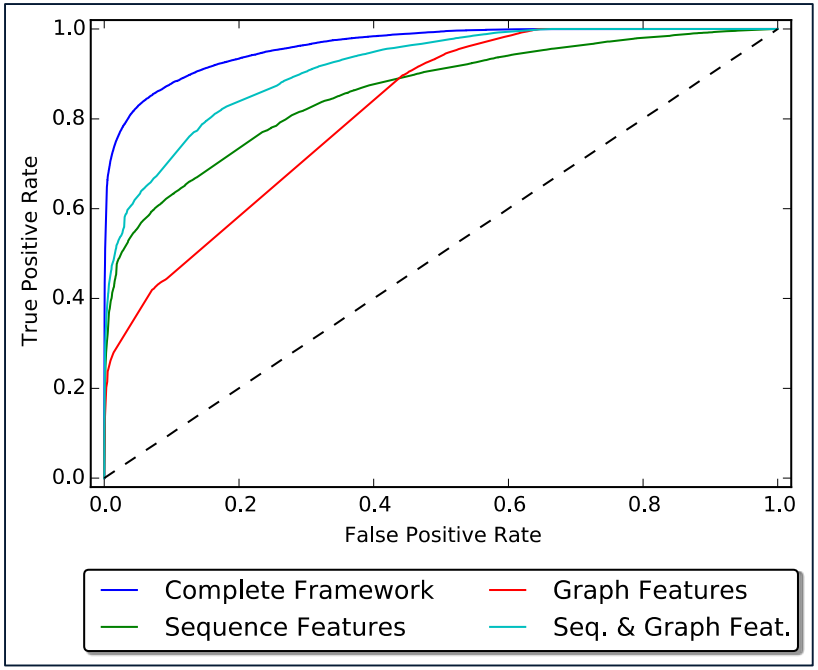
Little benefit from MMM (although little overhead)

Results

Precision-Recall



ROC



We can classify 70% of the spammers that need manual labeling with about 90% accuracy

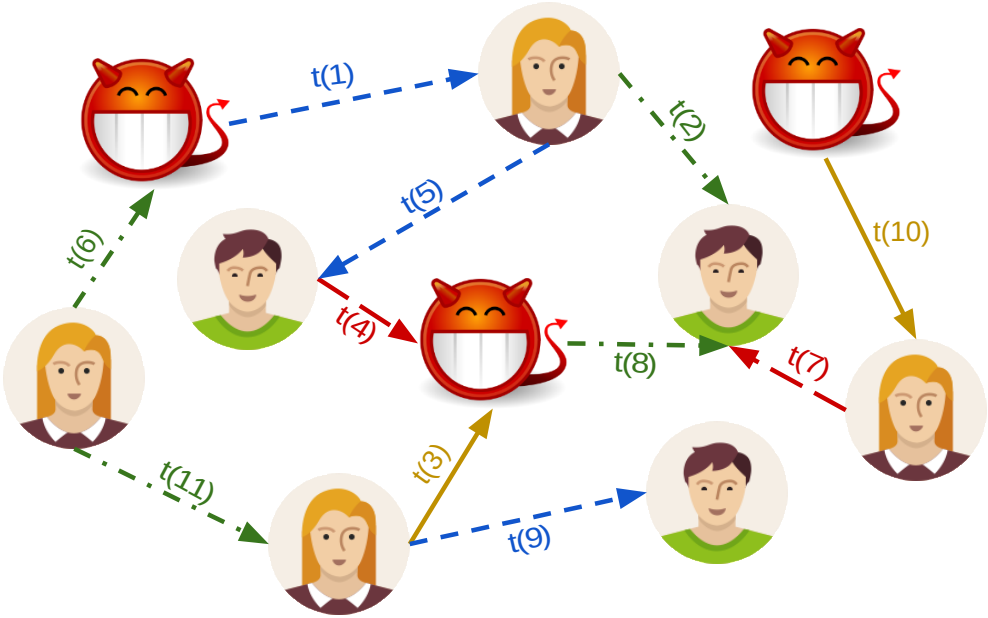
Deployment and Example Runtimes

- 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:

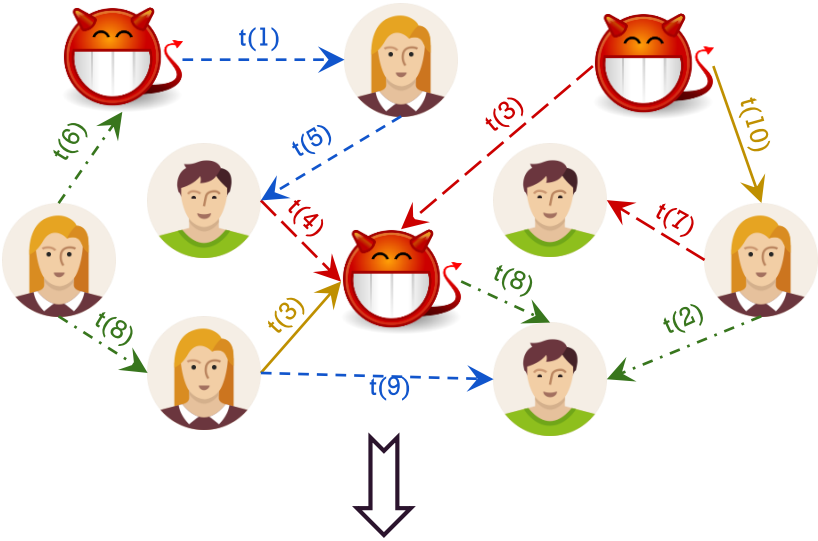
- Graph structure
- Action sequences
- **Reporting behavior**

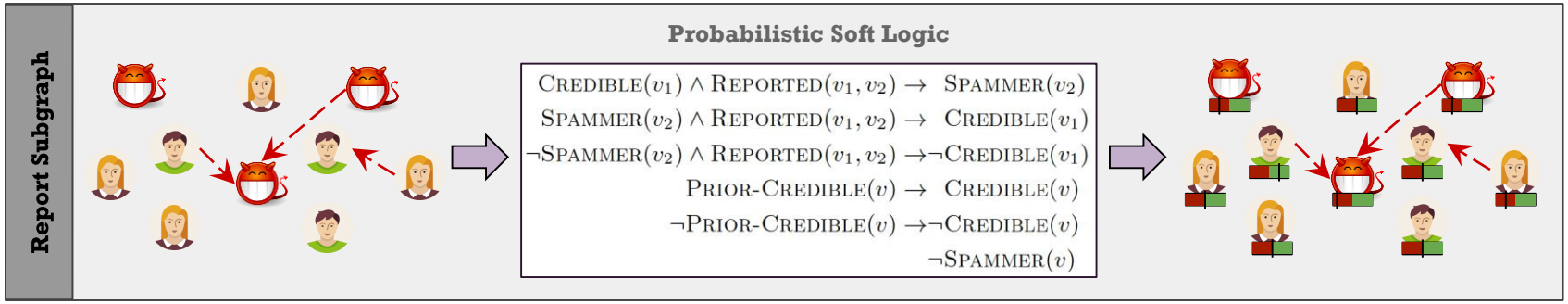


Refining the abuse reporting systems

- 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





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:

$$\omega : P(A, B) \wedge Q(B, C) \rightarrow R(A, C)$$

- 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:

$$REPORTED(v_1, v_2) \rightarrow \quad SPAMMER(v_2)$$
$$\neg SPAMMER(v)$$

Collective Classification with Reports

- Model using reports and credibility of the reporter:

$$
\begin{aligned}
CREDIBLE(v_1) \wedge REPORTED(v_1, v_2) &\rightarrow \quad SPAMMER(v_2) \\
PRIOR\text{-}CREDIBLE(v) &\rightarrow \quad CREDIBLE(v) \\
\neg PRIOR\text{-}CREDIBLE(v) &\rightarrow \neg CREDIBLE(v) \\
&\quad \neg SPAMMER(v)
\end{aligned}
$$

Collective Classification with Reports

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

$$
\begin{aligned}
CREDIBLE(v_1) \wedge REPORTED(v_1, v_2) &\rightarrow \quad SPAMMER(v_2) \\
SPAMMER(v_2) \wedge REPORTED(v_1, v_2) &\rightarrow \quad CREDIBLE(v_1) \\
\neg SPAMMER(v_2) \wedge REPORTED(v_1, v_2) &\rightarrow \neg CREDIBLE(v_1) \\
PRIOR\text{-}CREDIBLE(v) &\rightarrow \quad CREDIBLE(v) \\
\neg PRIOR\text{-}CREDIBLE(v) &\rightarrow \neg CREDIBLE(v) \\
&\quad\;\; \neg SPAMMER(v)
\end{aligned}
$$

Results of Classification Using Reports

Experiments	AU-PR	AU-ROC
Reports Only	0.674 ± 0.008	0.611 ± 0.007
Reports & Credibility	0.869 ± 0.006	0.862 ± 0.004
Reports & Credibility & Collective Reasoning	0.884 ± 0.005	0.873 ± 0.004

Results of Classification Using Reports

Experiments	AU-PR	AU-ROC
Reports Only	**0.674 ± 0.008**	**0.611 ± 0.007**
Reports & Credibility	0.869 ± 0.006	0.862 ± 0.004
Reports & Credibility & Collective Reasoning	0.884 ± 0.005	0.873 ± 0.004

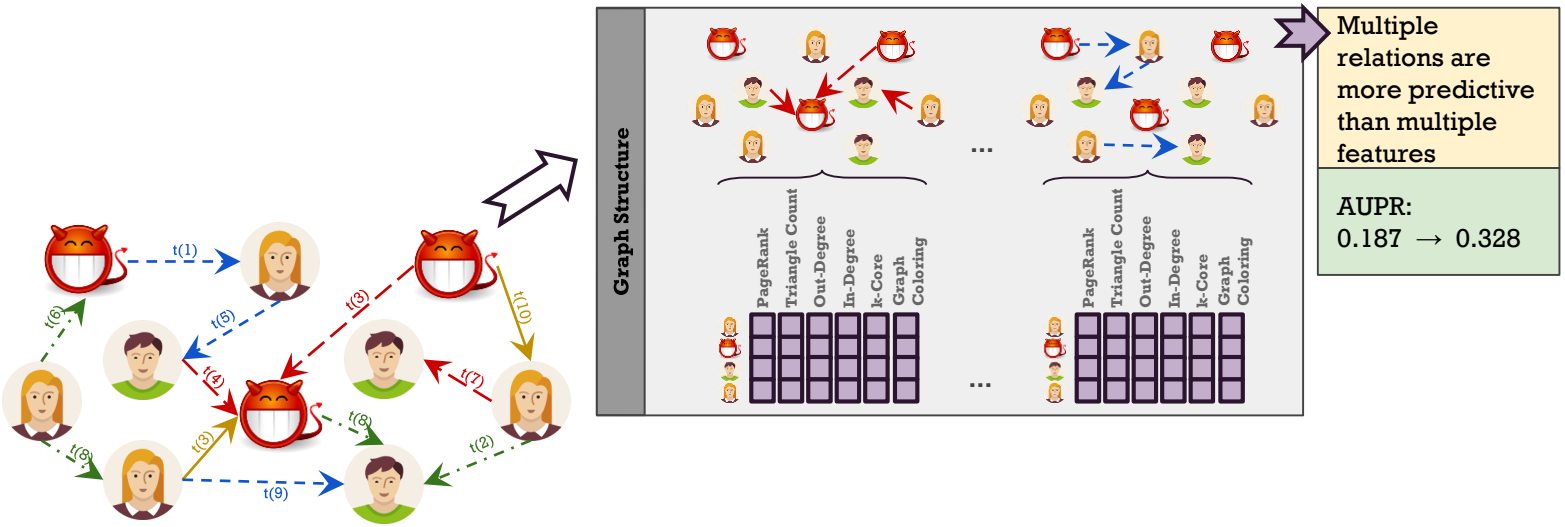
Results of Classification Using Reports

Experiments	AU-PR	AU-ROC
Reports Only	0.674 ± 0.008	0.611 ± 0.007
Reports & Credibility	**0.869 ± 0.006**	**0.862 ± 0.004**
Reports & Credibility & Collective Reasoning	0.884 ± 0.005	0.873 ± 0.004

Results of Classification Using Reports

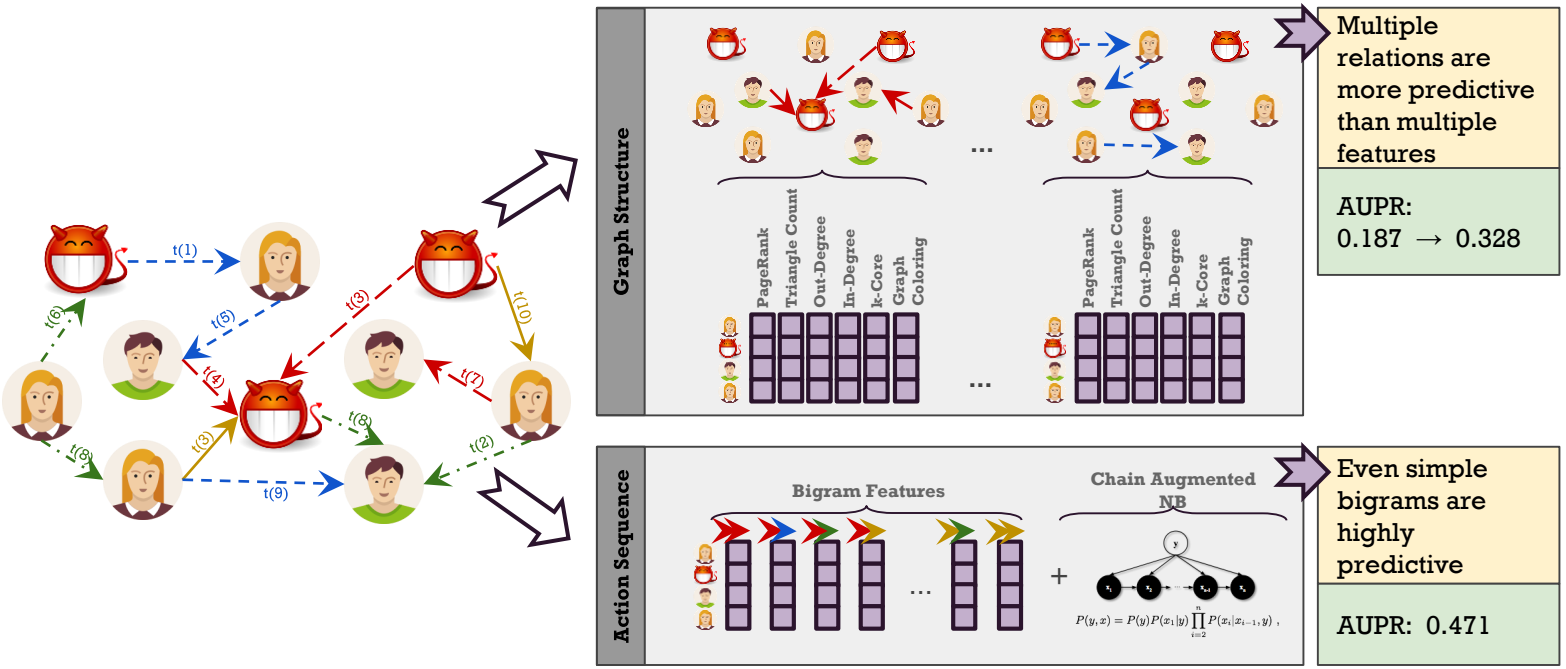
Experiments	AU-PR	AU-ROC
Reports Only	0.674 ± 0.008	0.611 ± 0.007
Reports & Credibility	0.869 ± 0.006	0.862 ± 0.004
Reports & Credibility & Collective Reasoning	**0.884 ± 0.005**	**0.873 ± 0.004**

Conclusion



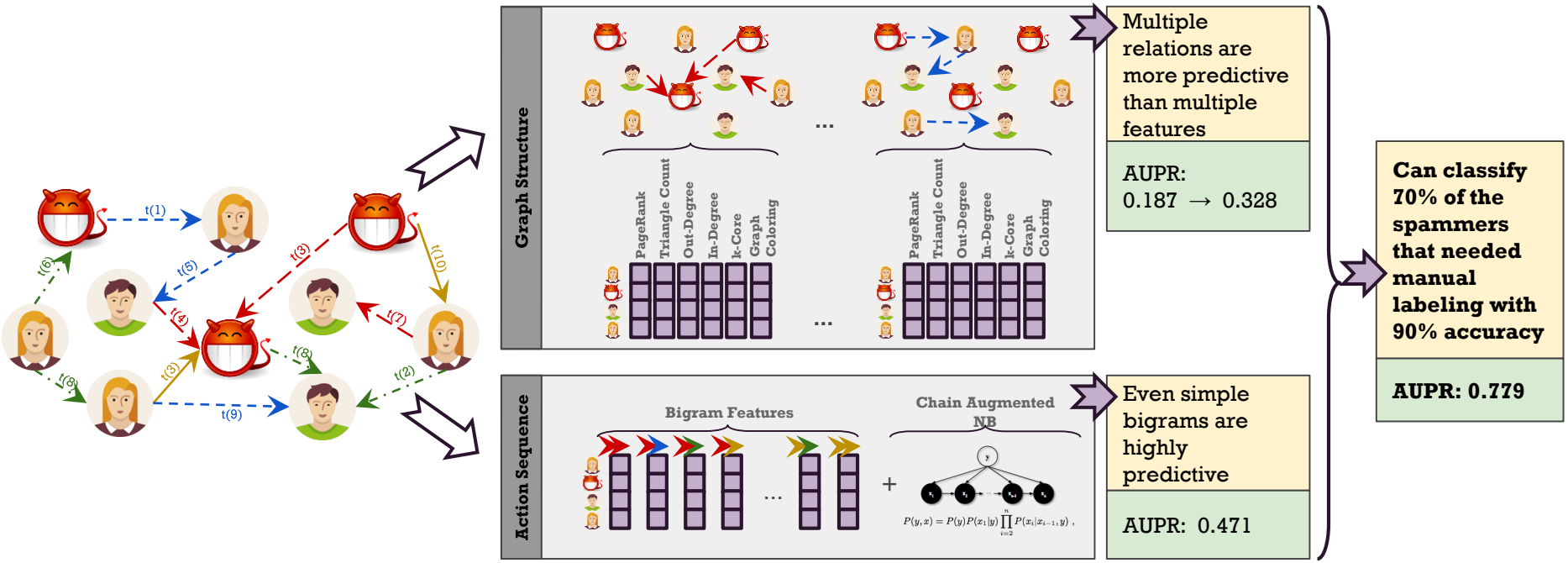
Code and part of the data will be released soon:
https://github.com/shobeir/fakhraei_kdd2015

Conclusion



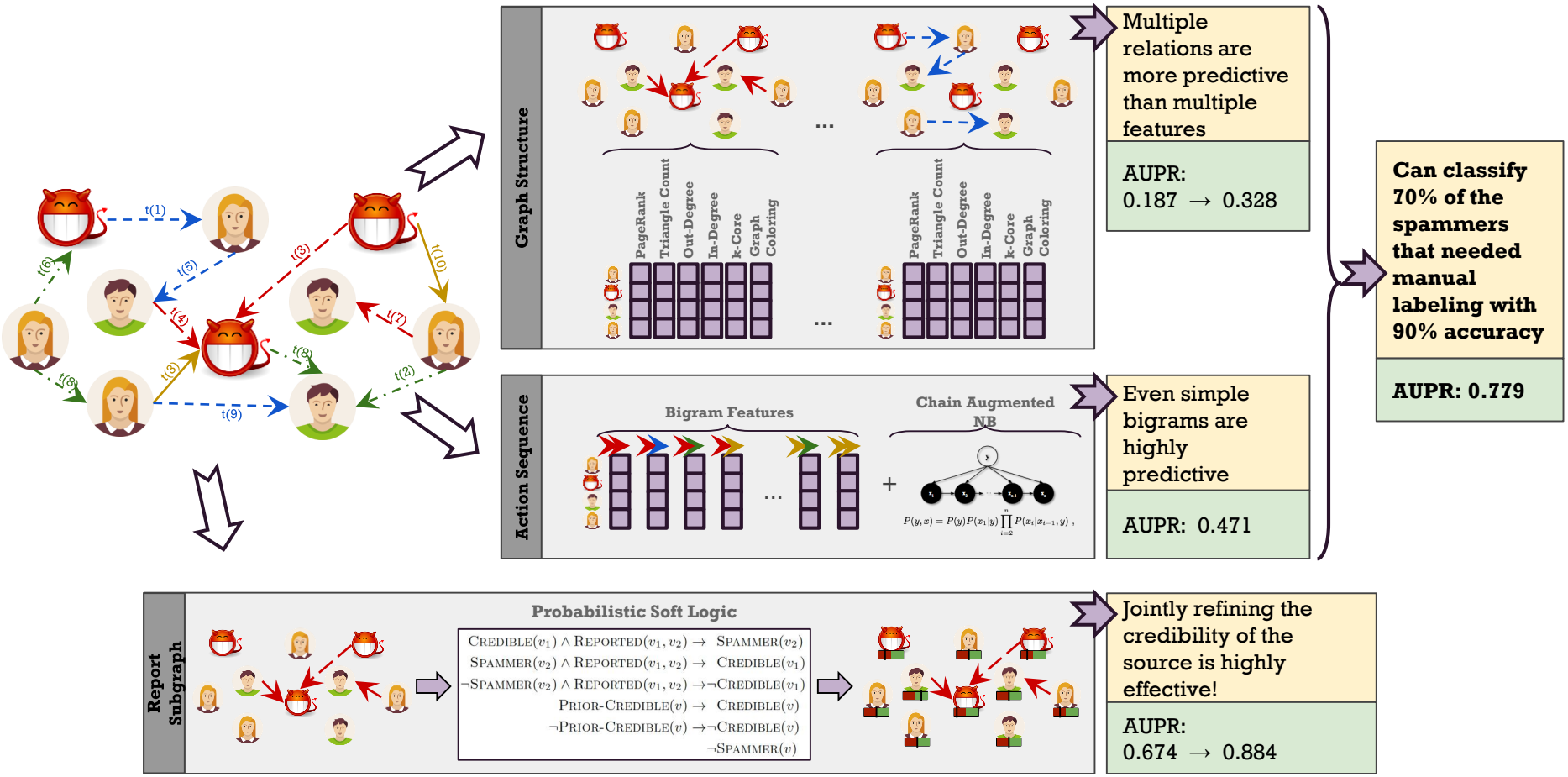
Code and part of the data will be released soon:
https://github.com/shobeir/fakhraei_kdd2015

Conclusion



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

Conclusion



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

Acknowledgements

- Collaborators:

Lise Getoor
Univ. California, Santa Cruz

Shobeir Fakhraei
Univ. of Maryland

Madhusudana Shashanka
if(we) Inc., currently Niara Inc.

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

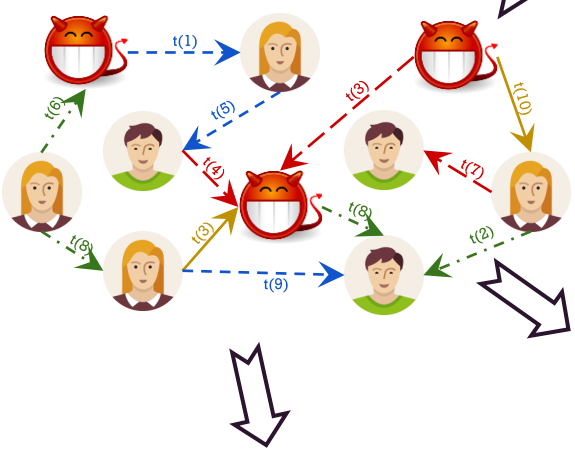
Johann Schleier-Smith, Karl Dawson, Dai Li, Stuart Robinson, Vinit Garg, and Simon Hill

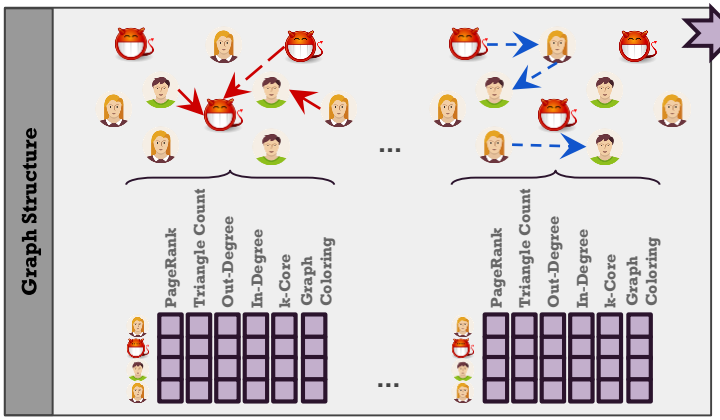
- Dato (Formerly Graphlab):

Danny Bickson, Brian Kent, Srikrishna Sridhar, Rajat Arya, Shawn Scully, and Alice Zheng

Conclusion

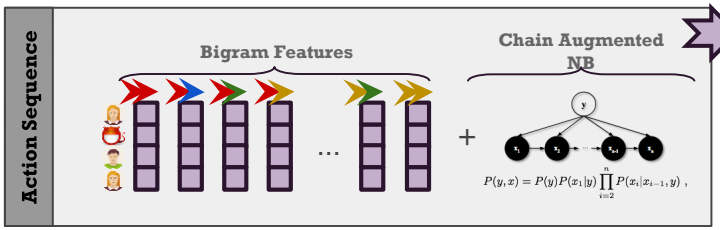
Thank you!





Multiple relations are more predictive than multiple features

AUPR:
0.187 → 0.328

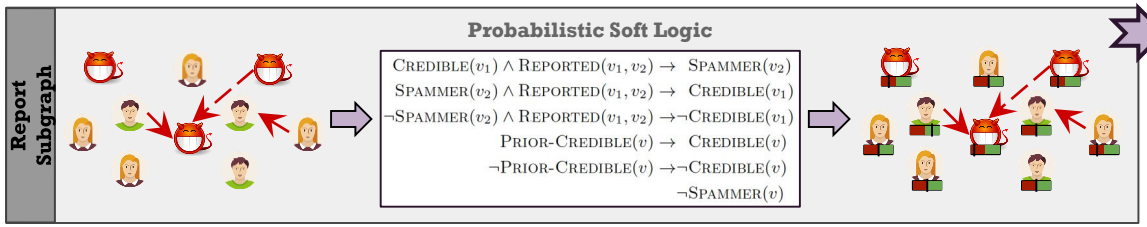


Even simple bigrams are highly predictive

AUPR: 0.471

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

AUPR: 0.779



Jointly refining the credibility of the source is highly effective!

AUPR:
0.674 → 0.884

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