Spam Classification Based on Signed Network Analysis

Sihyun Jeong and Kyu-haeng Lee

모바일시스템공학과
32181928 박찬호

Contents

Conventional Spam Classifications

- Spam detection = classification problem
- Identifying the appropriate data features to distinguish spam/non-spam
    - text examining : e.g. template-matching scheme
    - meta information (user profile, activity logs, timestamps)

→ become rapidly ineffective by cleverer and more multifaceted attacks

Conventional
Spam Classifications

- Network Properties - highlighted as alternative features for spam classification

  - not easily imitated by attackers

- Thus, social network analysis are more robust (Node Ranking; PageRank, HITS)
- However, attackers adroitly avoid such spam detection

Proposed Scheme

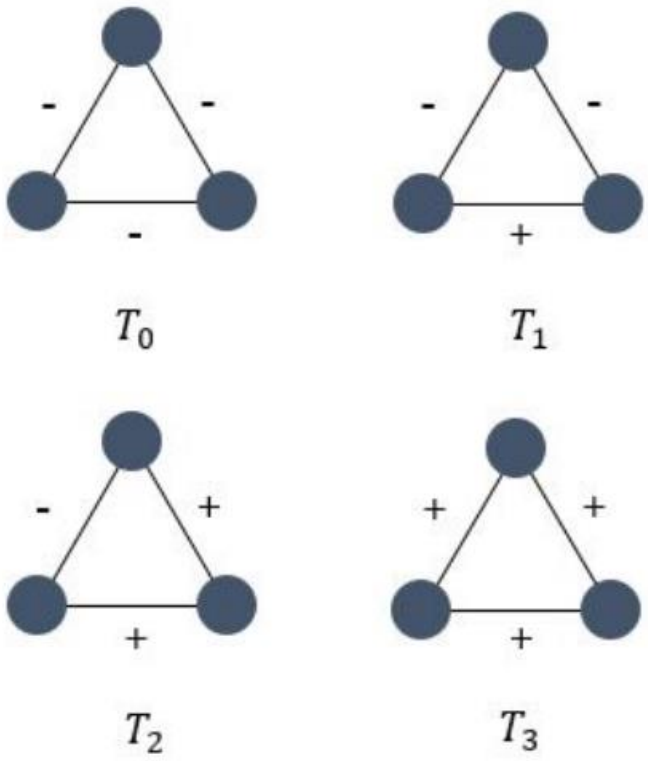
- Similar approaches with proposed scheme
    - NFS (Network Footprint Score)
    - Graph propagation algorithm for fake accounts detection (Li et al.)
    - Integro
    - SybilEdge
    - Triad Significance Profile (TSP)-Filtering
- limitation - restricted to only undirected or unsigned network

Proposed Scheme

- **Structural Balance Theory (= Balance Theory)**
- **Social Status Theory (= Status Theory)**
- **Concept of Surprise**
- **Graph-Converting Method**

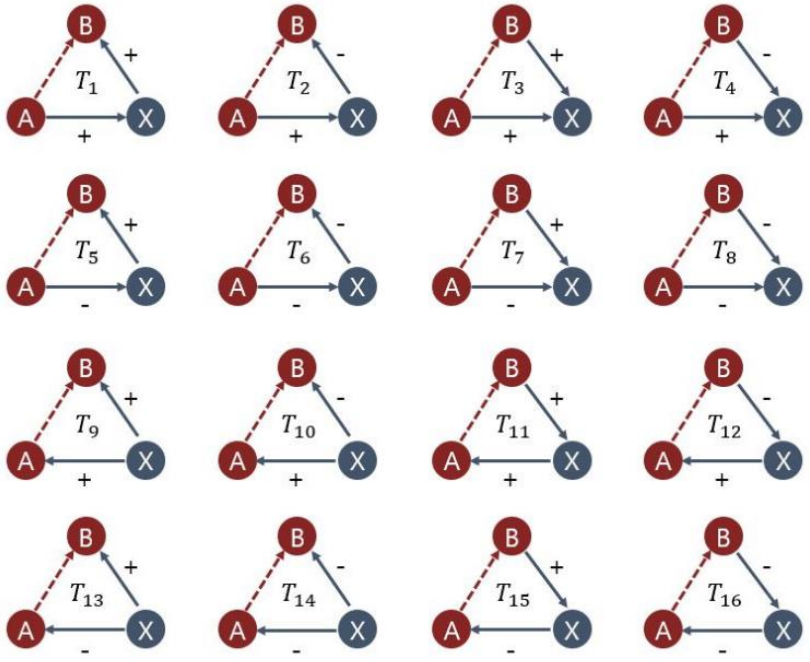
Proposed Scheme
- Balance Theory

- Balanced (plausible) : T_1 , T_3 // Unbalanced (not plausible) : T_0 , T_2
- The friend of my friend is my friend (T_1)
- The enemy of my friend is my enemy (T_3)
- Intended for undirected graphs
- a decrease in classification performance due to the information loss for edge directionality



Proposed Scheme
- Status Theory

- determined by the status difference between nodes
- a positive(negative) directed link : the recipient as having higher(lower) status
- Sixteen triad types for signed directed networks
- predicting the sign of the red link considering the signs of the A–X and B–



Concept of Surprise

- certain type of motif
- shows how much the target motif appears in the actual network compared to our expectation

Concept of Surprise (Balance Theory)

- T_i : number of certain triad T_i
- $E[T_i]$: expected number of type T_i
- Δ : the total number of triads in the network
- $\Delta p_0(T_i)$: expected fraction of triads that are of type T_i
- $s_b(T_i)$: surprise value for a certain motif T_i

$$s_b(T_i) = \frac{T_i - E[T_i]}{\sqrt{\Delta p_0(T_i)(1 - p_0(T_i))}}$$

Concept of Surprise (Status Theory)

- Measure surprise separately for the edge initiator and the edge destination
- Generative baseline, Receptive baseline
- computed by the number of standard deviations by which the actual number of positive A–B edges in the data differs from the expected positive numbers created by the baseline

Graph-Converting Method

- Among all E edges, we select α · E edges and assign them negative signs
- α is the fraction of negative edges (0.2 from real world data)
- how to appropriately select α · E edges

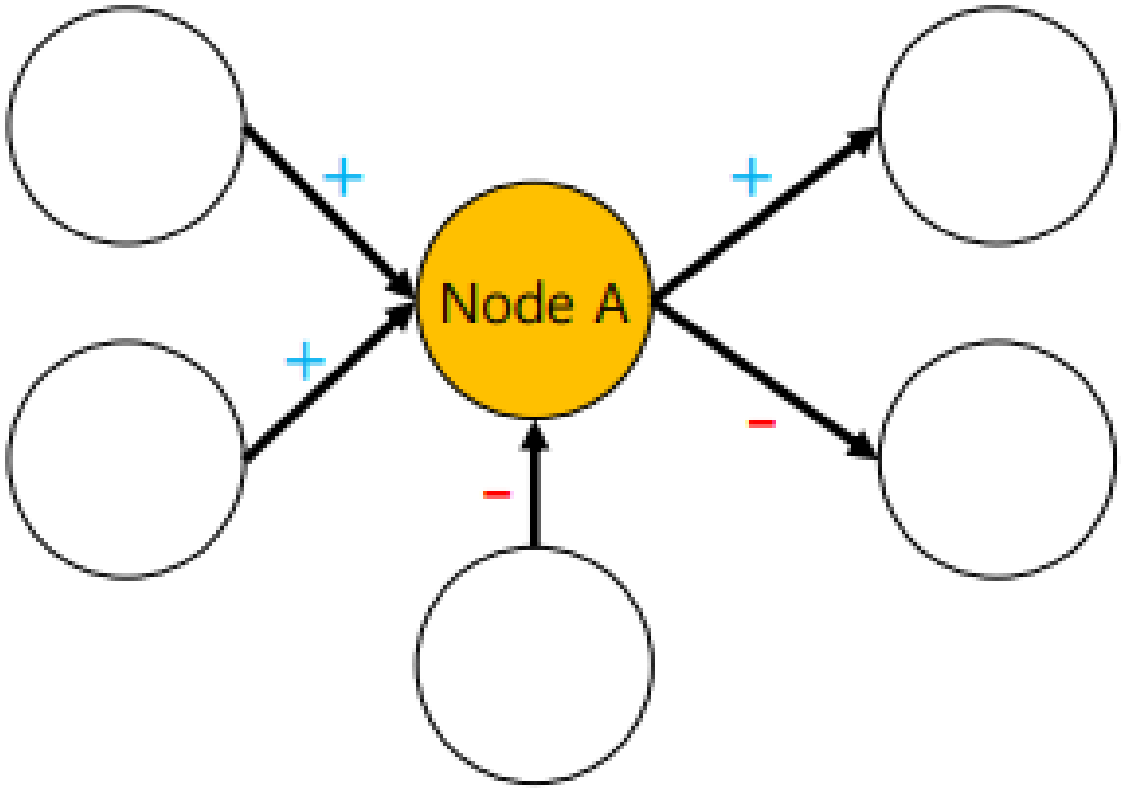
Table 1. Statistics of the Twitter dataset.

	Twitter
The number of users	54,981,152
The number of spammers	41,352
The number of following links	1,963,263,821

Table 2. Statistics of the Epinions dataset.

	Epinions
The number of nodes	131,828
The number of edges	841,372
The ratio of positive edges	85%
The ratio of negative edges	15%

PageRank

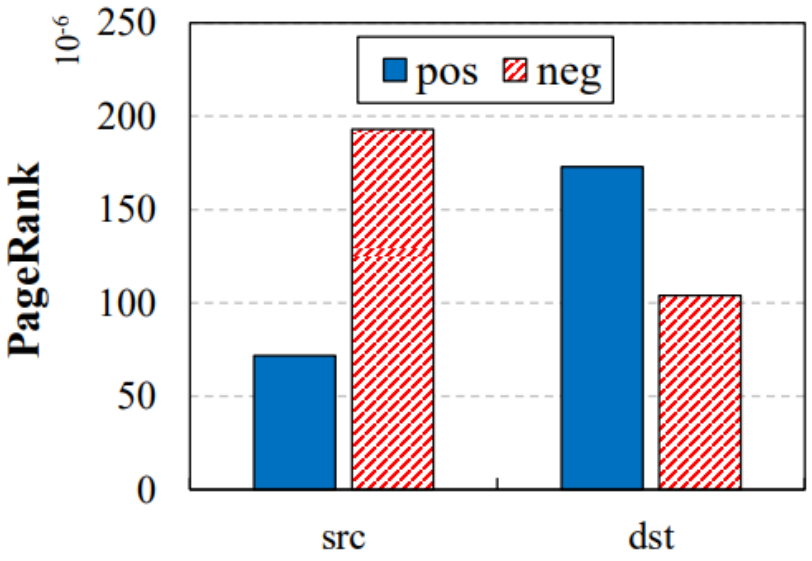


Positive : ratio of incoming = 2/3

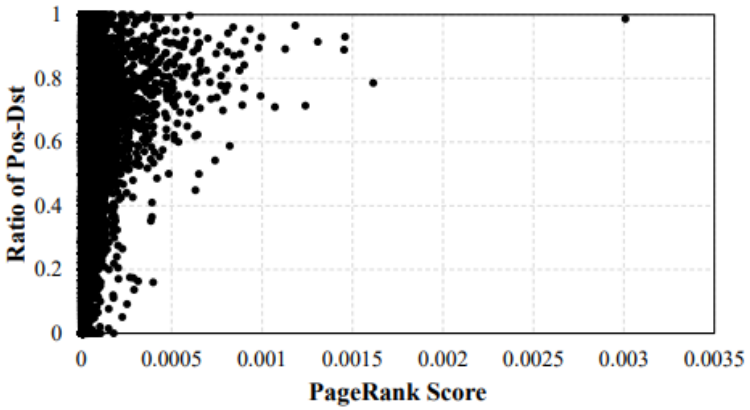
Negative : ratio of incoming = 1/2

PageRank

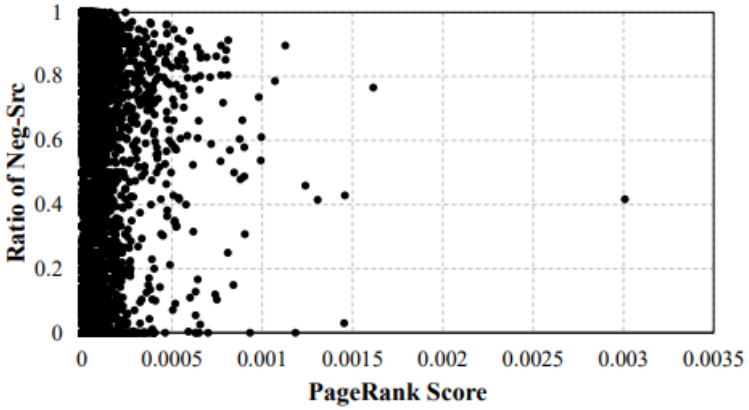
high ratio values of positive incoming edges → high PageRank scores



(c)Average PageRank score.



(a)Ratio of positive incoming edges.



(b)Ratio of negative outgoing edges.

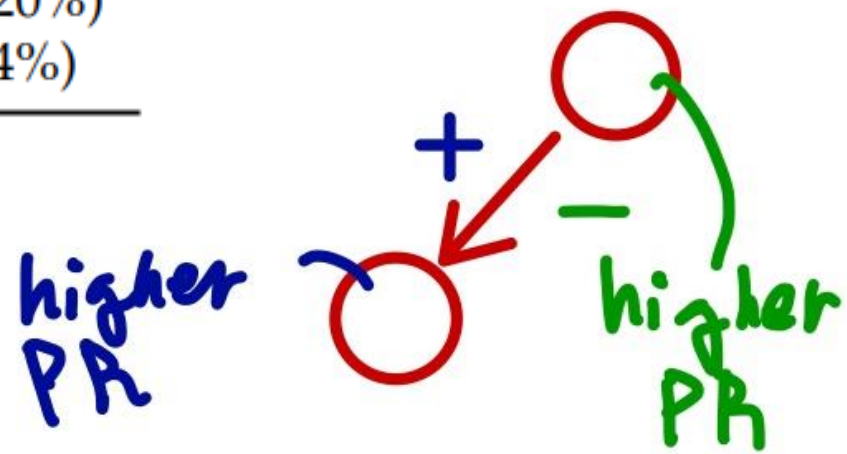
PageRank

PR(x) = PageRank score of node x, dst = destination node, src = source node

d **= PR(dst) – PR(src)**

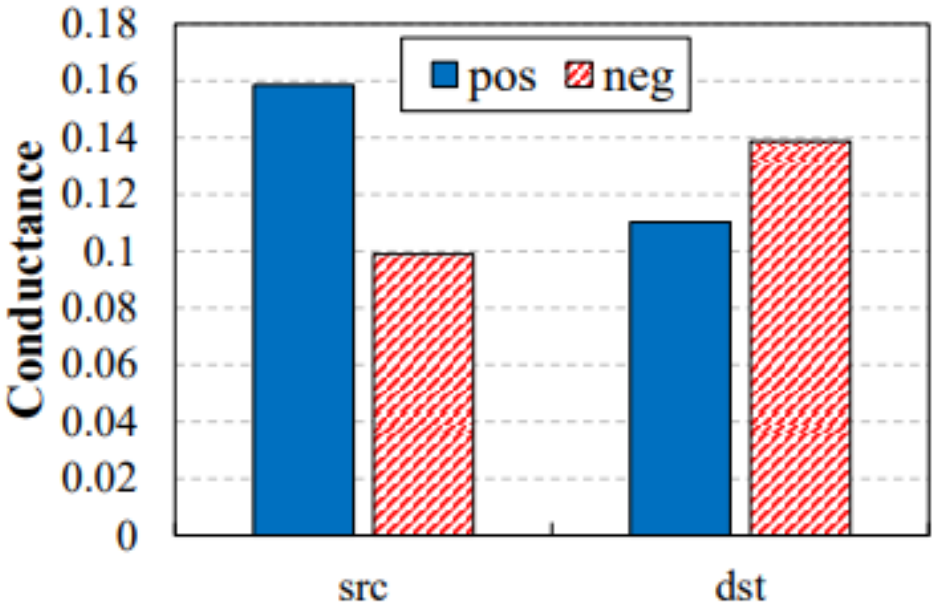
Table 3. The number of edges according to PageRank differences.

	$d < 0$	$d > 0$	$d = 0$
Positive Edge	464,119 (64.67%)	252,135 (35.13%)	1412 (0.20%)
Negative Edge	43,261 (34.97%)	80,396 (64.99%)	49 (0.04%)



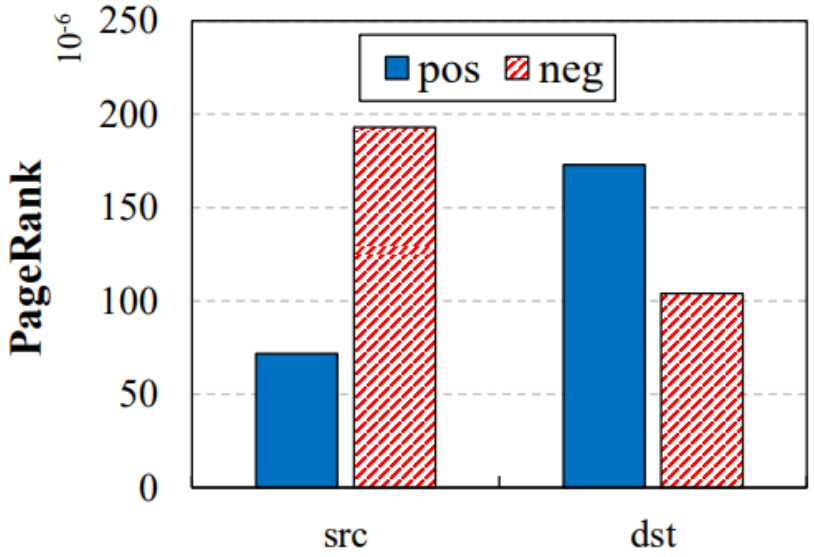
Conductance

- generate clusters → measure the Conductance for each cluster that an edge belongs to
- internal and external connectivity
- higher Conductance means higher separability of the cluster
- From the analysis, conductance score is
  - Positive edges : src node > dst node
  - Negative edges : src node < dst node
- ↑ internal connectivity & ↓ external separability
- = ↓ conductance → ↑ PageRank score

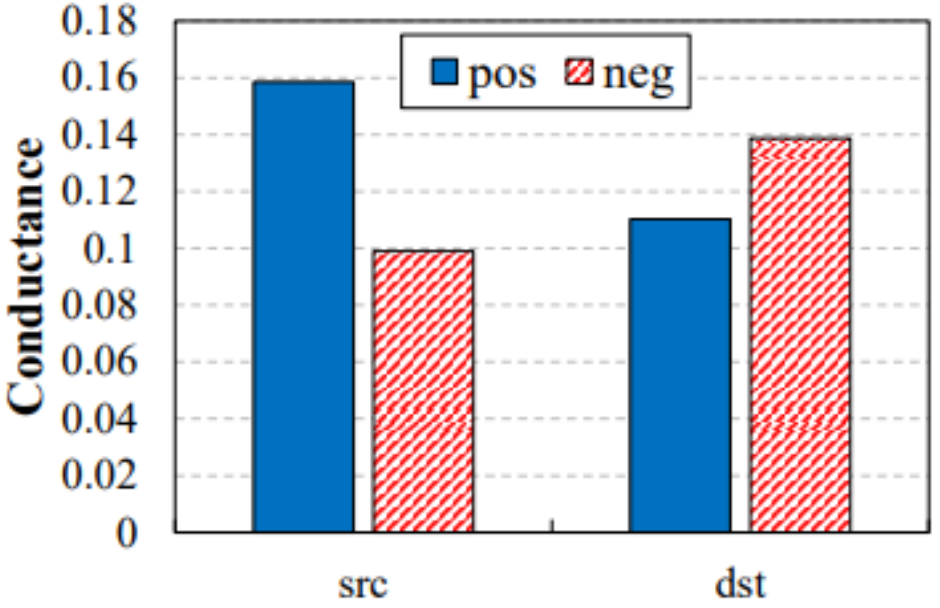


Conductance

- generate clusters → measure the Conductance for each c
- internal and external connectivity
- higher Conductance means higher separability of the clu
- From the analysis, conductance score is
  - Positive edges : src node > dst node
  - Negative edges : src node < dst node
- ↑ internal connectivity & ↓ external separability
- = ↓ conductance → ↑ PageRank score



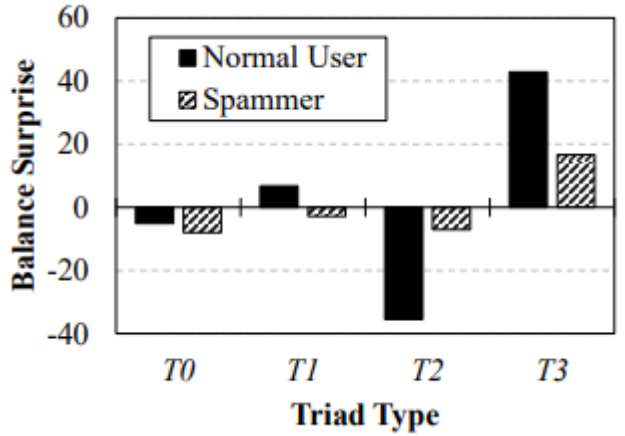
(c)Average PageRank score.



Validation

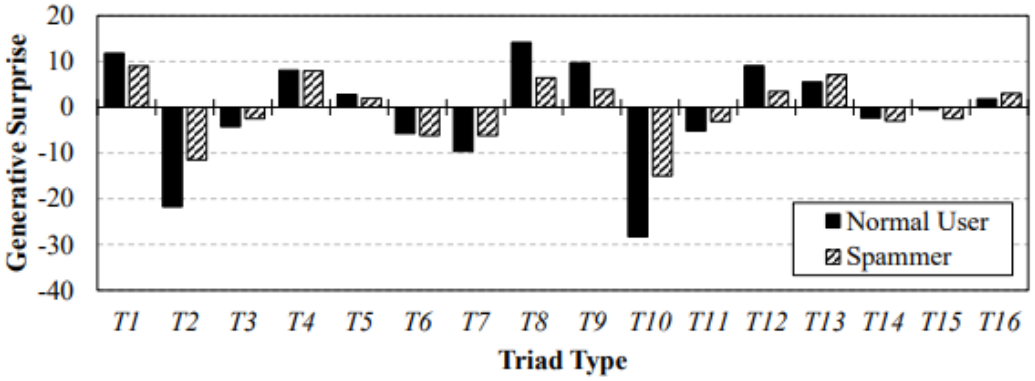
Balance

- Spammers are inconsistent with T_1
- T_2 , T_3 → spammers' gap is smaller

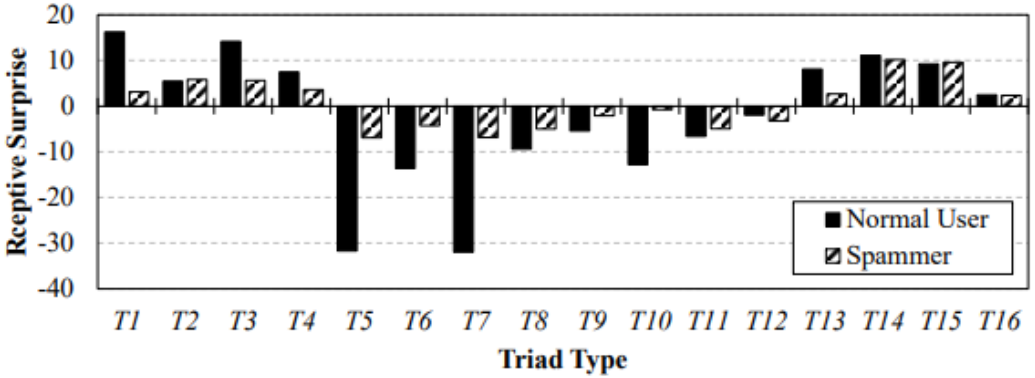


Status

- Spammers are consistent with different degrees
- Generative : T_2 , T_8 , T_{10}
- Receptive : most types of triads



(a)Generative surprise



(b)Receptive surprise

Performance Evaluation

- Proposed method
  - Best accuracy, precision
  - Recall of CatchSync, TSP is higher

    with higher false positive rate(FPR)
  - Best performance when only using Statu

Table 5. Performance comparison result.

	Accuracy	Precision	Recall	FPR	F1-Score
SybilRank	0.283	0.303	0.335	0.769	0.318
NFS	0.572	0.597	0.443	0.299	0.509
CatchSync	0.903	0.894	0.915	0.109	0.904
TSP	0.908	0.906	0.911	0.095	0.908
Balance	0.925	0.943	0.904	0.055	0.923
Status	0.929	0.949	0.907	0.049	0.927
Balance + Status	0.928	0.948	0.906	0.050	0.926

Contribution

- First spam classification utilizing structural balance theory and social status theory
  - the edge signs are highly likely to be determined by considering users' social relationships
  - a substantial difference between the edge sign patterns of spammers and those of non-spammers
- Graph-converting method using PageRank and Conductance scores

Future Work

- **Recommendation System – the sign of an edge might be determined based on ratings**
- **Sentiment analysis**
- **Advanced Classifiers with graph-embedding methods**
- **Semi-supervised learning for given network data lack information (sign)**

Thank You

감사합니다.