A Survey and Critique of Recent Advance in Online Spam Detection

Yingtong Dou

Department of Computer Science
University of Illinois at Chicago

March 9, 2019

Road Map

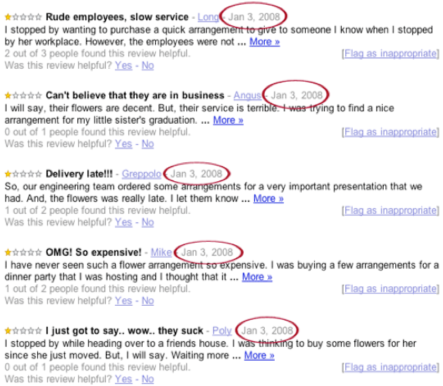
What is spam?

- Spam is the **fake** and **useless** information on the Internet.
- Spam is composed by **intentionally crafted** content.
- Spam is everywhere.

Examples of Spams

Twitter

Amazon

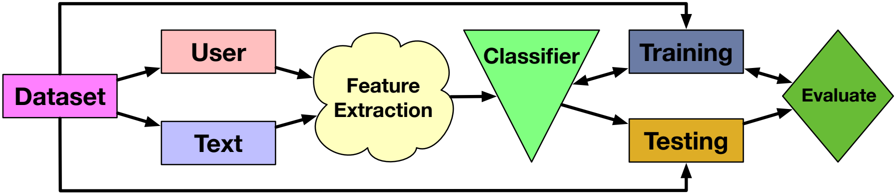


Effect of Spams

- Decay online experience.
- Bias users' choice/opinion.
- Mislead recommender system.
- Usually come with other security threats.

Spam Detection Problem

- Spam detection is an **Anomaly Detection** task in data mining.
- We need to evaluate the **suspiciousness** of users, posts, reviews.
- Generally, spam detection is a **Supervised Learning** task.



Feature Extraction

- Semantic Features

Feature Name	Description
RL	Average review length
ACS	Average content similarity
PCW	Percentage of all capital words
PC	Percentage of capital letters
DL_b	Description length based on bigrams
PP1	The ratio of 1st person pronouns
RES	The ratio of exclamation sentences
SW	The ratio of subjective words
OW	The ratio of objective words
F	The frequency of review

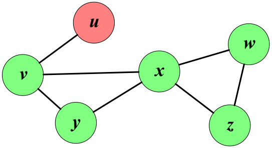
Feature Extraction

- Behavioral Features

Behavior	Description
MNR	Max. number of reviews posted in a day
PR	The ratio of positive reviews
NR	The ratio of negative reviews
WRD	Weighted Rating Deviation
BST	Burstiness
RD	Rating deviation of product's avg. rating
Rank	The rank order of the review
ETF	The early time frame of the reviewer
ISR	Is singleton?
DPW	Deceptive review count previous week

Learning Methods

- **Traditional Model**
  Naive Bayes, Support Vector Machine, Random Forest etc.
- **Graph Model**



- **Deep Model**
  Relation Embedding, Convolutional Neural Network etc.

Overview of the Three Paper

Paper	Problem	Model	Target	Dataset	Venue
Paper 1	Optimizing Alg	Graph	Social Spammer	Twitter	ICDM2017
Paper 2	Cold Start	Deep Model	Spam Review	Yelp	ACL2017
Paper 3	Crowdsourcing	Graph&Deep	Crowd Worker	Amazon	WSDM2018

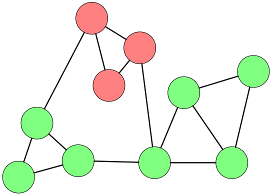
GANG: Detecting Fraudulent Users in Online Social Networks via Guilt-by-Association on Directed Graphs

Binghui Wang, Neil Zhenqiang Gong
ECE Department, Iowa State University
{binghuiw, neilgong}@iastate.edu

Hao Fu
Microsoft Research Asia, China
fuha@microsoft.com

Graph Homophily Assumption

- Similar nodes are more likely to connect with each other than dissimilar ones.
- In the online social network, we represent **users as nodes**, their **friendship as edges**.
- **Suspicious users** tend to connect with each other; **regular users** tend to connect with each other.



Modeling Directed Edge Influence

Bidirectional Edge

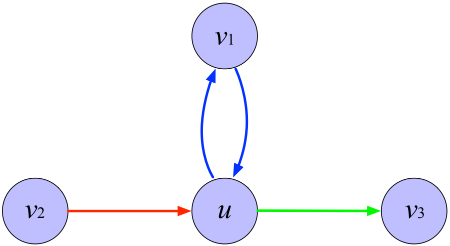
$v_1(Benign) \Rightarrow u(Benign)$
$v_1(Suspicious) \Rightarrow u(Suspicious)$

Unidirectional Incoming Edge

$v_2(Benign) \Rightarrow u(Benign)$
$v_2(Suspicious) \Rightarrow u(?)$

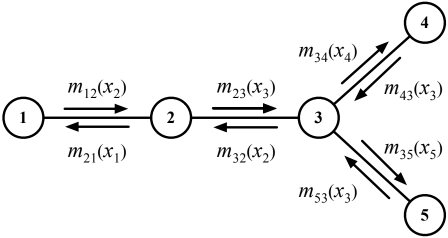
Unidirectional Outgoing Edge

$v_3(Benign) \Rightarrow u(?)$
$v_3(Suspicious) \Rightarrow u(Suspicious)$



Inference on MRF

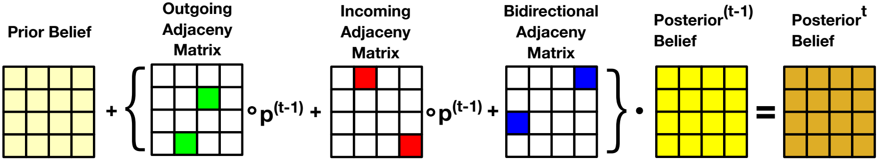
- The social network could be modeled as a Markov Random Field.
- Belief propagation could infer the states of the nodes in MRF.



Message Passing Process of Belief Propagation

Optimizing Belief Propagation

- Eliminate message maintenance in BP.
- Approximate BP using matrix multiplication.



One Round of Approximated BP in Matrix Form

Contributions of the Paper

- Models the influence of directed edges with a unified formulation.
- Proves the convergence condition of the proposed model.
- Strong theoretical guarantee.
- Algorithm is scalable .

Drawbacks of the Paper

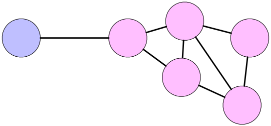
- Model is vulnerable to attacks.
- It is a trade off between model robustness and model performance

Handling Cold-Start Problem in Review Spam Detection by Jointly Embedding Texts and Behaviors

Xuepeng Wang[1,2], **Kang Liu**[1], and **Jun Zhao**[1,2]
[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
{xpwang, kliu, jzhao}@nlpr.ia.ac.cn

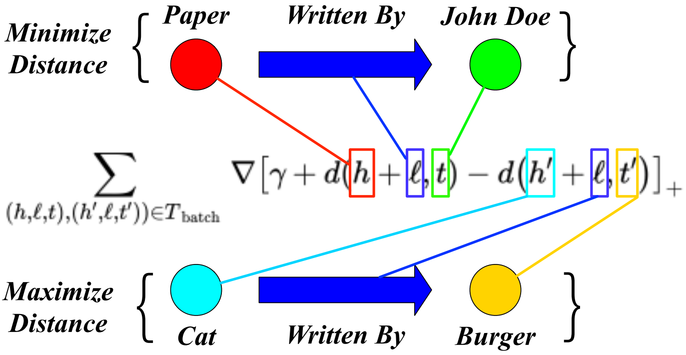
Cold Start Problem

- Cold Start refers to those new coming data items.
- Traditional features fail to model new users and reviews.
- The graph model is not useful for dealing with new users.



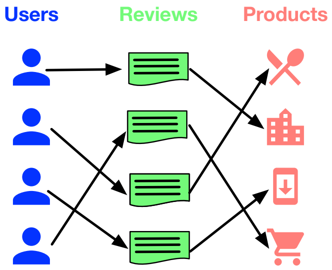
Translating Model

- Translating is a relation embedding model on knowledge graph.



Review Triplet

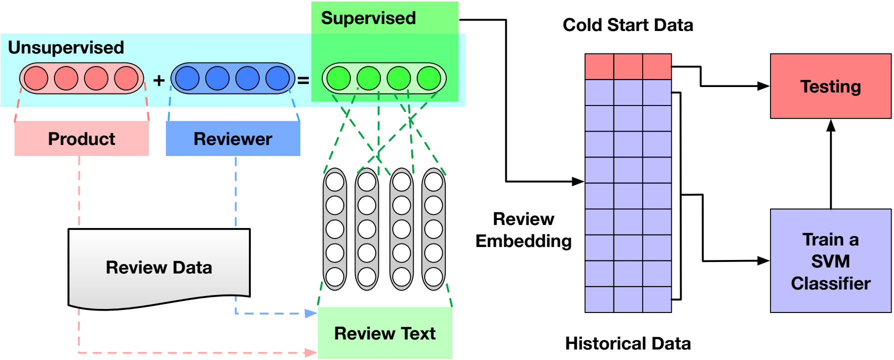
- Each review could be represented as a user-review-product triplet.



Relations in a Review Platform

Review Relation Embedding

- Product as head, reviewers as relation, review embedding as tail.



Contributions

- First paper tackling cold start in spam detection.
- Encode latent relations with deep network.

Drawbacks

- No explanation for review triplet setting.
- No comparison with other embedding models.
- No comparison with other dimension reduction models.

Combating Crowdsourced Review Manipulators: A Neighborhood-Based Approach

Parisa Kaghazgaran
Texas A& M University
College Station, TX
kaghazgaran@tamu.edu

James Caverlee
Texas A& M University
College Station, TX
caverlee@tamu.edu

Anna Squicciarini
Pennsylvania State University
State College, PA
asquicciarini@ist.psu.edu

Crowdsourcing Spam

- Crowdsourcing is an activity that hires online freelancing workers to finish specific tasks.

Available Tasks	Amount	Time
Youtube: Vote for this video	$0.10	1 min
Follow me on Twitter	$0.12	1 min
Insurance Form: Sign up	$1.50	5 min
Create Gmail account for me	$0.13	3 min
Online Game: Sign up	$0.20	3 min
Digg: Bookmark my page	$0.10	1 min
Upload 5 photos to this site	$0.39	4 min

A Screenshot of Crowdsourcing Tasks Listed on a Website

Challenges in Defending Crowdsourcing Attack

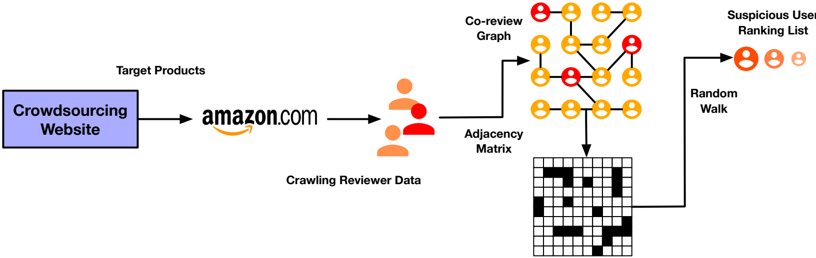
- Crowd workers look like regular users.
- It is difficult to acquire the ground truth.

Select Seed Users

- Crawl all the Amazon products that have released tasks on a crowdsourcing website.
- Crawl all the reviewers having reviewed those products.
- Select seed users.

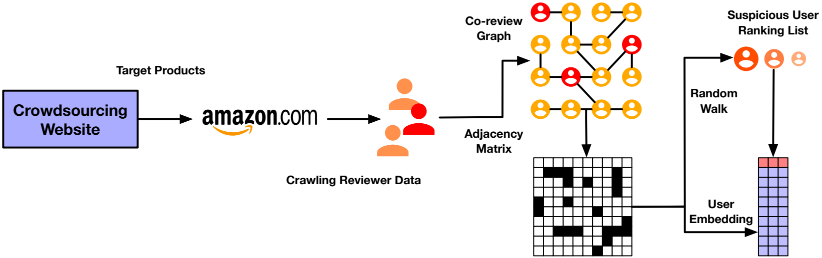
Discover Local Similar Users

- Construct a co-review graph where reviewers are nodes. Edges connect users who both have reviewed the same product.
- Set the suspicious score of seed users to 1.
- Using random walk to propagate the suspiciousness.



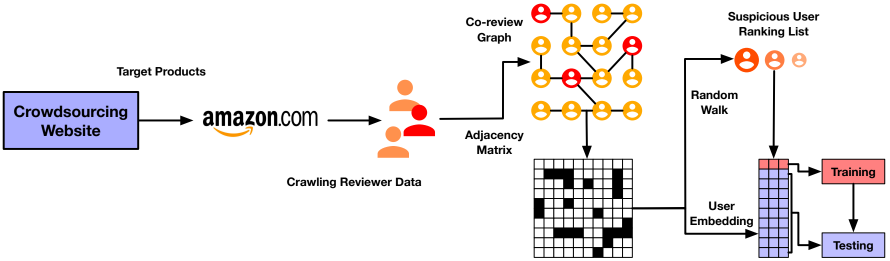
Discover Distant Similar Users

- Random walk only discovers local similar users. It cannot pass suspiciousness to users having no connection with seed users.
- Use **node2vec** model to learn the node embedding.
- Use node embedding to find structural similar users.



TwoFace Framework

- Select seed users; use random walk to generate suspicious user ranking list; learn node embeddings of all users.
- Train traditional classifiers with node embeddings.
- Validate the model with holdout data.



Model Limitations

- No side information of users.
- No comparison with other embedding models.

Summary of Three Papers

- Traditional features and models have limitations.
- Graph models have strong theoretical guarantee.
- Deep models have weak interpretability.

Challenges in Spam Detection Research

- Vulnerability to attacks.
- Reproducibility of deep model.
- Lack of theoretical guarantee.
- Quality of benchmark datasets.
- Practical performance of detectors.

Adapt More Deep Models

- Graph Convolutional Network.
- Long Short Term Memory Network.
- Auto-Encoder Framework.

Other Promising Research Directions

- Adversarial machine learning.
- Dynamic detection model.
- Heterogeneous information network.
- New problems:
  - Poisoning reviews
  - Fake news
  - Multi-intention reviews
  - Machine-generated content

Appendix I

Formulation of the proposed model in Paper 1.

$$
\begin{cases}
\mathbf{A}_i'^{(t-1)} = I\left(\mathbf{A}_i \circ \mathbf{P}^{(t-1)^T}\right) \\
\mathbf{A}_o'^{(t-1)} = I\left(-\mathbf{A}_o \circ \mathbf{P}^{(t-1)^T}\right) \\
\mathbf{p}^{(t)} = \mathbf{q} + 2 \cdot w \cdot \left(\mathbf{A}_b + \mathbf{A}_i'^{(t-1)} + \mathbf{A}_o'^{(t-1)}\right) \cdot \mathbf{p}^{(t-1)}
\end{cases}
$$

A_i: incoming edge adjacency matrix

A_o: outgoing edge adjacency matrix

A_b: bidirectional edge adjacency matrix

P: node posterior belief

q: node prior belief

w: coupling strength

Appendix II

Algorithm 1 Learning TransE

input Training set $S = \{(h, \ell, t)\}$, entities and rel. sets E and L, margin γ, embeddings dim. k.

initialize $\boldsymbol{\ell} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $\ell \in L$
        $\boldsymbol{\ell} \leftarrow \boldsymbol{\ell} / \|\boldsymbol{\ell}\|$ for each $\ell \in L$
        $\mathbf{e} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$
loop
    $\mathbf{e} \leftarrow \mathbf{e} / \|\mathbf{e}\|$ for each entity $e \in E$
    $S_{batch} \leftarrow \text{sample}(S, b)$ // sample a minibatch of size b
    $T_{batch} \leftarrow \emptyset$ // initialize the set of pairs of triplets
    for $(h, \ell, t) \in S_{batch}$ **do**
        $(h', \ell, t') \leftarrow \text{sample}(S'_{(h,\ell,t)})$ // sample a corrupted triplet
        $T_{batch} \leftarrow T_{batch} \cup \{((h, \ell, t), (h', \ell, t'))\}$
    end for
    Update embeddings w.r.t. $\displaystyle\sum_{((h,\ell,t),(h',\ell,t')) \in T_{batch}} \nabla \left[\gamma + d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{\ell}, \boldsymbol{t'})\right]_+$
end loop

Reference

1 Wang, Binghui, Neil Zhenqiang Gong, and Hao Fu. "GANG: Detecting fraudulent users in online social networks via guilt-by-association on directed graphs." In 2017 IEEE International Conference on Data Mining (ICDM), pp. 465-474. IEEE, 2017.

2 Wang, Xuepeng, Kang Liu, and Jun Zhao. "Handling cold-start problem in review spam detection by jointly embedding texts and behaviors." In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 366-376. 2017.

3 Kaghazgaran, Parisa, James Caverlee, and Anna Squicciarini. "Combating crowdsourced review manipulators: A neighborhood-based approach." In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 306-314. ACM, 2018.

4 Rayana, Shebuti, and Leman Akoglu. "Collective opinion spam detection: Bridging review networks and metadata." In Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining, pp. 985-994. ACM, 2015.

5 Mukherjee, Arjun, Vivek Venkataraman, Bing Liu, and Natalie Glance. "What yelp fake review filter might be doing?." In Seventh international AAAI conference on weblogs and social media. 2013.