Two Step Graph-based Semi-supervised Learning for Online Auction Fraud Detection

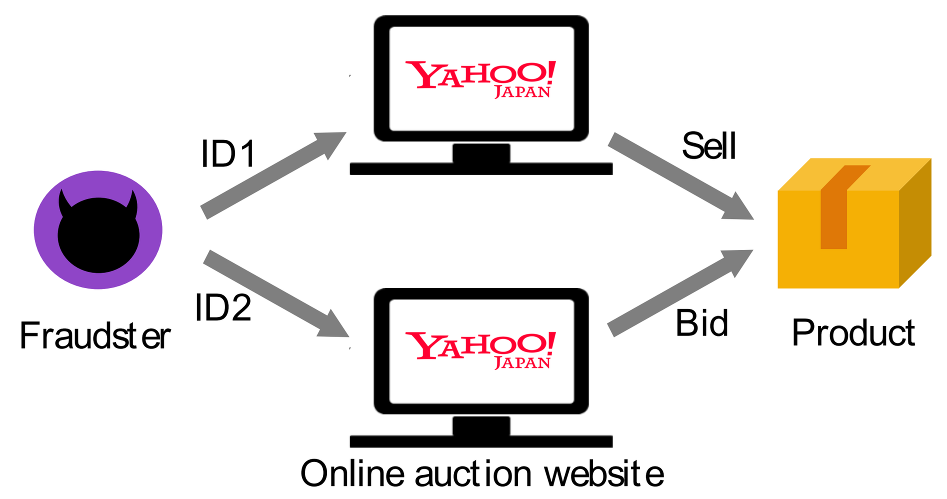
European Conference on Machine Learning and Principles and Practice of Knowledge Discovery (ECML-PKDD 2015)

Phiradet Bangcharoensap[1], Hayato Kobayashi[2], Nobuyuki Shimizu[2], Satoshi Yamauchi[2], and Tsuyoshi Murata[1]

[1]Tokyo Institute of Technology, [2]Yahoo Japan Corporation

Definition of Fraudster

Auction users who participate in their own auction with other user IDs in order to drive up the final price.

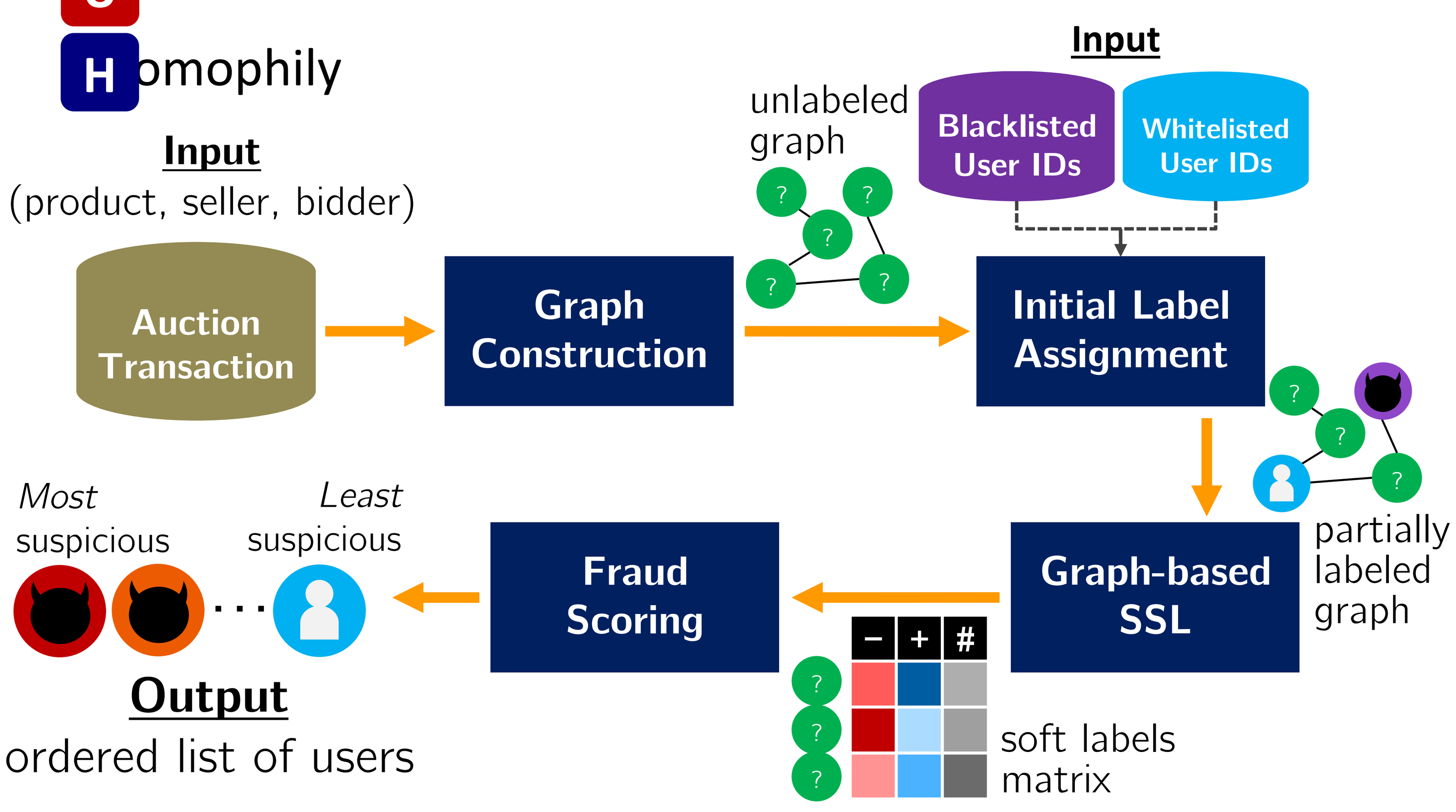


Overview

Key Ideas

U niform interaction of innocents

H omophily



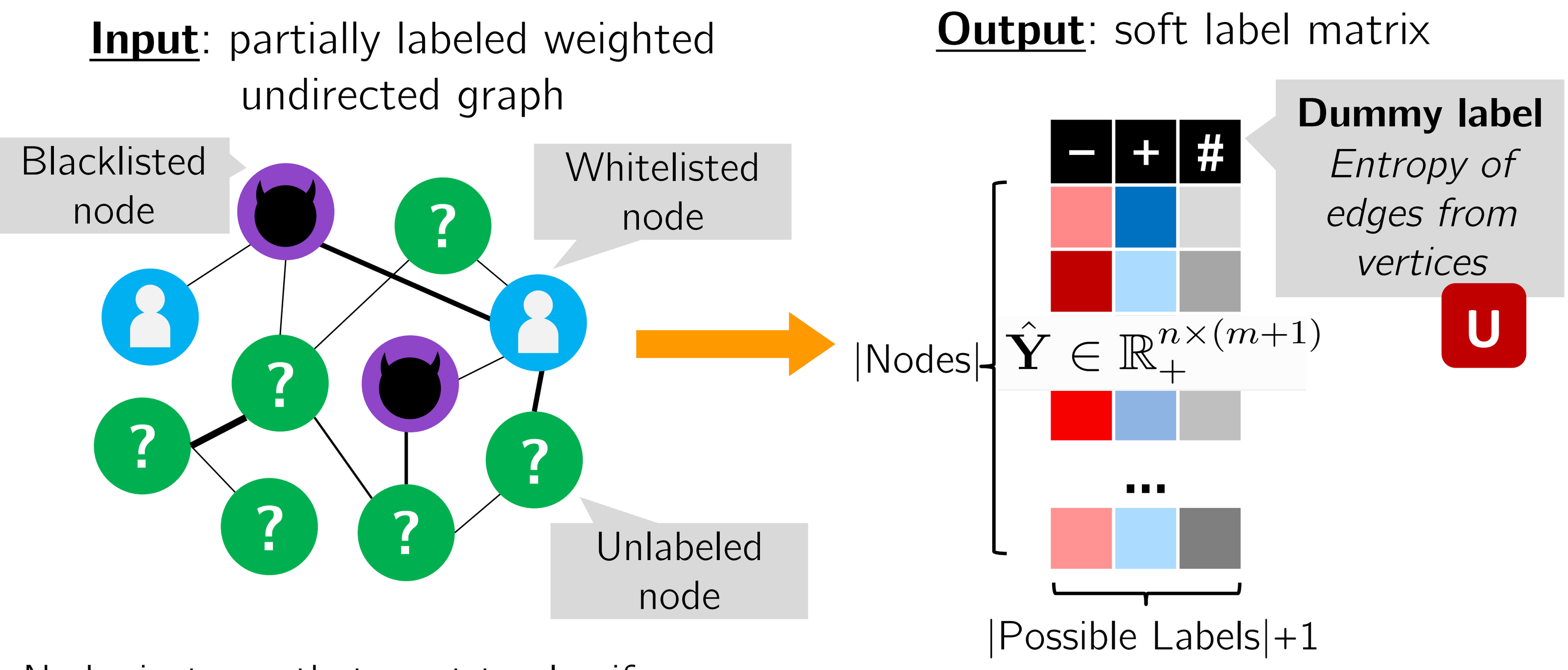
Objective: Fraudsters working in the same collusion with blacklisted users are ranked at the top.

Graph-base SSL

Modified Adsorption (MAD) [1]

Input: partially labeled weighted undirected graph

Output: soft label matrix



Node: instance that want to classify

- A few labelled nodes, with confidence

Edge: similarity between instances

Assign a score indicating likelihood of being each label (soft labels)

Tradeoff between *fitting* and *smoothness* constraints

- ***Fitting***: retain initial labels of seed nodes
- ***Smoothness:*** assign same labels to adjacent nodes **H**

Solve the convex optimization problem

$$\min_{\hat{\mathbf{Y}}} \sum_{l\in\mathcal{L}} \left[\mu_1(\mathbf{Y}_l - \hat{\mathbf{Y}}_l)^\top \mathbf{S}(\mathbf{Y}_l - \hat{\mathbf{Y}}_l) + \mu_2 \hat{\mathbf{Y}}_l^\top \mathbf{L} \hat{\mathbf{Y}}_l + \mu_3 \left\| \hat{\mathbf{Y}}_l - \mathbf{R}_l \right\|^2 \right]$$

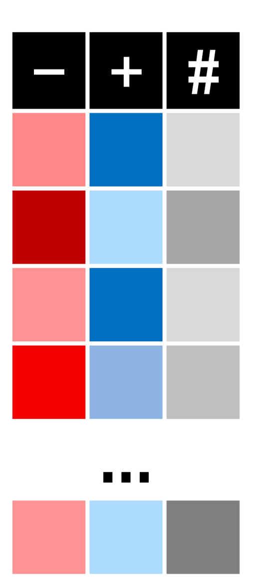
Fitting, Smoothness, Regularization

$\hat{\mathbf{Y}}$ where is a matrix storing score of each label, **Y** stores seed information, and **L** is the Laplacian matrix.

Fraud Scoring

Input

soft label matrix



Bad, Good, Dummy

Output

fraud score of nodes

MAD

$$\varphi(v, \hat{\mathbf{Y}}) = \frac{\hat{\mathbf{Y}}_{v1}}{\sum_{l=1}^{m+1} \hat{\mathbf{Y}}_{vl}}$$

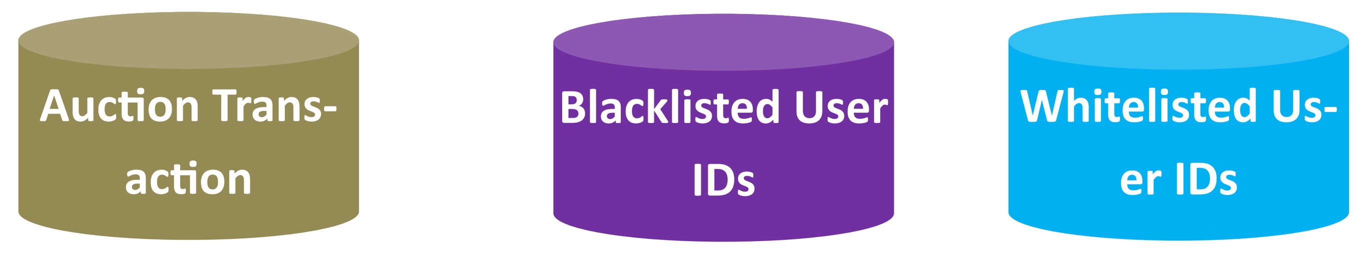
2-STEP

$$\rho(v, \hat{\mathbf{Y}}) = \varphi(v, \hat{\mathbf{Y}}) + \frac{\gamma}{|N(v)|} \sum_{u\in N(v)} \mathbf{W}_{uv}\, \varphi(u, \hat{\mathbf{Y}})$$

Weighted degree centrality (WDC)

Experiment Setting

Real-world Dataset



(product, seller, bidder)
≈ 16 million transactions
≈ 2 million users

≈ 550 users

Official store account
≈ 10,000 users

Evaluation Metric

Higher NDCG is better.

The predicted ranking results were compared with the blacklisted users. We used normalized discounted cumulative gain (NDCG) [2] as the evaluation metric.

$$\text{NDCG} = \frac{\text{DCG}}{\text{IDCG}} \quad \text{DCG} = \sum_{i=1}^{p} \frac{2^{r(i)} - 1}{\log_2(i+1)} \quad \text{IDCG} = \sum_{i=1}^{\min(p,|Q|)} \frac{1}{\log_2(i+1)},$$

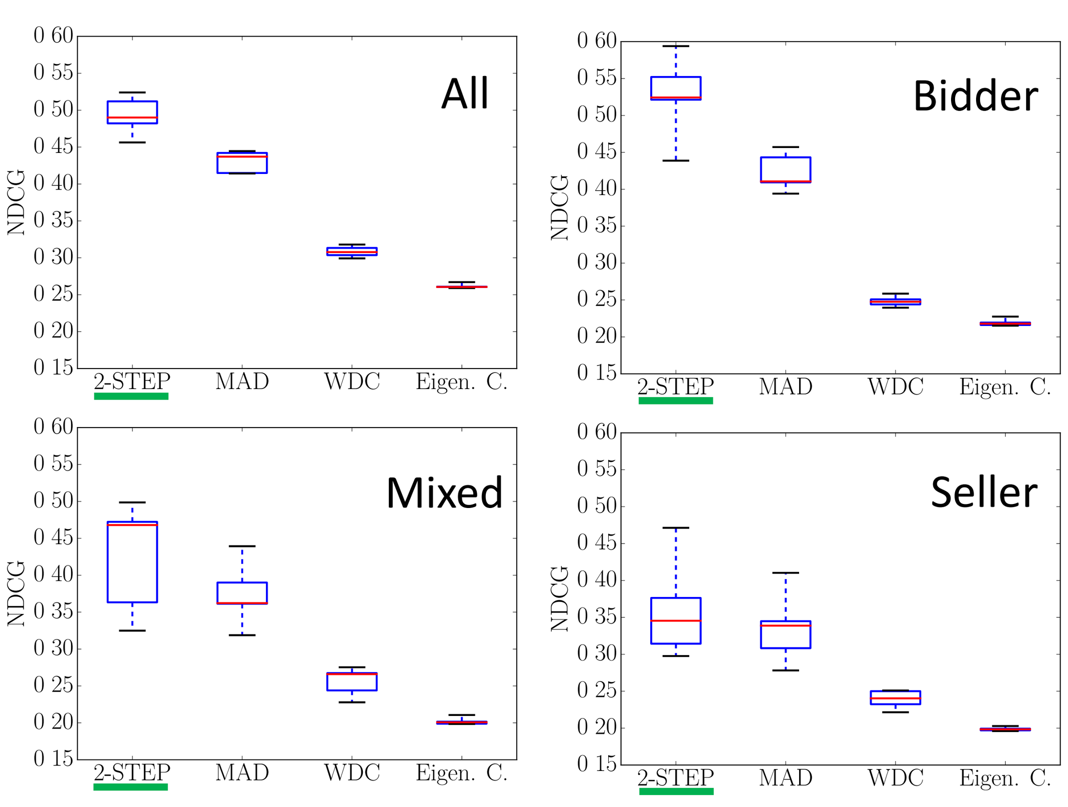
where p is the size of the result, $r(i)$ is one if the i^{th} result is fraudulent, and $|Q|$ is the number of testing fraudsters.

Result

Does the dummy label help?

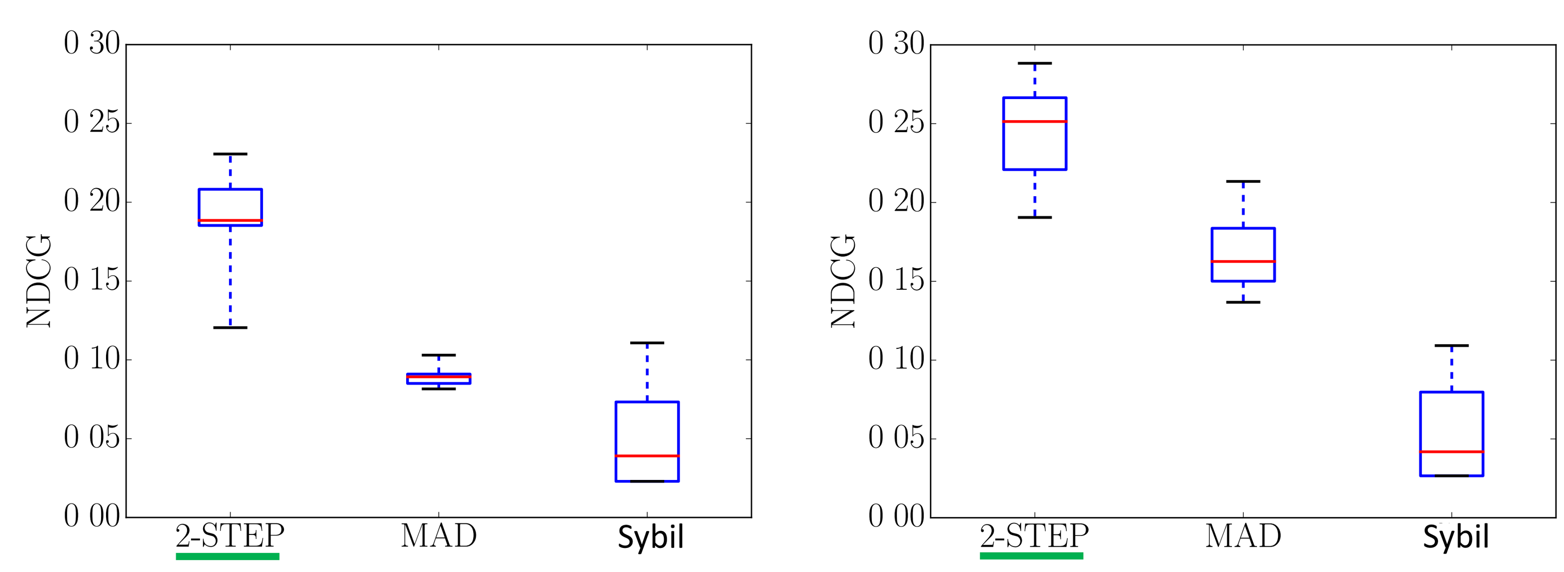
Node type	with dummy		w/o dummy	
	<NDCG>	SD	<NDCG>	SD
All	0.431	0.015	0.406	0.019
Bidder	0.423	0.026	0.397	0.035
Seller	0.336	0.049	0.284	0.029
Mixed	0.374	0.044	0.319	0.024

Comparison with unsupervised methods

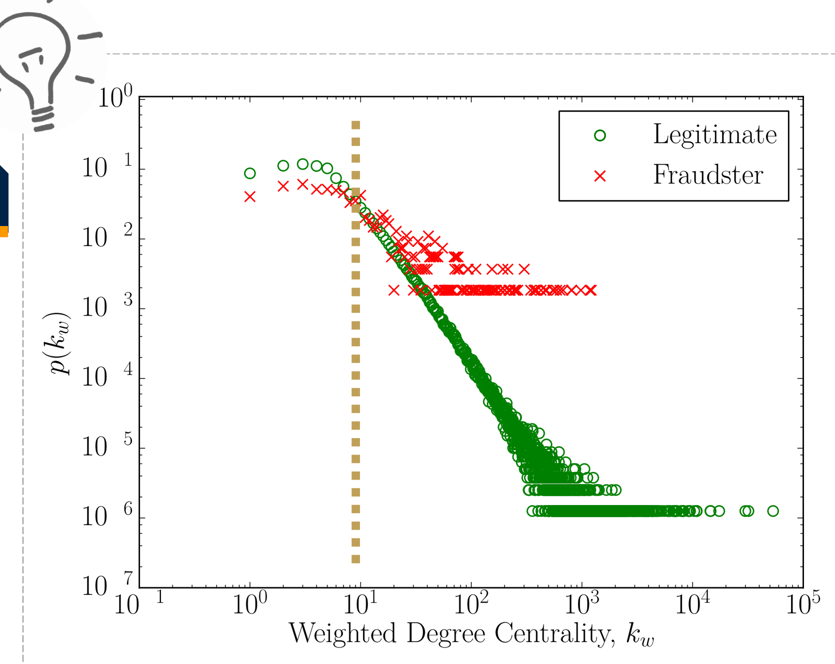


- Compare with weighted degree centrality (WDC) and eigenvector centrality (Eigen. C.)
- 2-STEP method outperforms MAD, WDC,

Comparison with a Sybil defense method [3]



Calculated from top 100 — Calculated from top 500



Distribution of weighted degree centrality (WDC) of legitimate users and fraudsters. Fraudsters tend to have higher WDC.

References

[1] Talukdar, P.P., Crammer, K.: New regularized algorithms for transductive learning. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II. pp. 442–457. ECML PKDD '09, Springer-Verlag, Berlin, Heidelberg (2009)

[2] Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 41–48. SIGIR'00, ACM, New York, NY, USA (2000)

[3] Viswanath, B., Post, A., Gummadi, K.P., Mislove, A.: An analysis of social network-based sybil defenses. SIGCOMM Computer Communication Review. 40(4), 363–374 (2010)