12-2012

Detecting Anomalies in Bipartite Graphs with Mutual Dependency Principles

Hanbo DAI
Singapore Management University, hanbo.dai.2008@smu.edu.sg

Feida ZHU
Singapore Management University, fdzhu@smu.edu.sg

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

Hwee Hwa PANG
Singapore Management University, hhpang@smu.edu.sg

Citation

Detecting Anomalies in Bipartite Graphs with Mutual Dependency Principles

Hanbo Dai   Feida Zhu   Ee-Peng Lim   HweeHwa Pang

School of Information Systems, Singapore Management University

{hanbo.dai.2008, fdzhu, eplim, hhpang}@smu.edu.sg

Abstract—**Bipartite graphs can model many real life applications including users-rating-products in online marketplaces, users-clicking-webpages on the World Wide Web and users-referring-users in social networks. In these graphs, the anomalousness of nodes in one partite often depends on that of their connected nodes in the other partite. Previous studies have shown that this dependency can be positive (the anomalousness of a node in one partite increases or decreases along with that of its connected nodes in the other partite) or negative (the anomalousness of a node in one partite rises or falls in opposite direction to that of its connected nodes in the other partite).**

In this paper, we unify both positive and negative mutual dependency relationships in an unsupervised framework for detecting anomalous nodes in bipartite graphs. This is the first work that integrates both mutual dependency principles to model the complete set of anomalous behaviors of nodes that cannot be identified by either principle alone. We formulate our principles and design an iterative algorithm to simultaneously compute the anomaly scores of nodes in both partites. Moreover, we mathematically prove that the ranking of nodes by anomaly scores in each partite converges. Our framework is examined on synthetic graphs and the results show that our model outperforms existing models with only positive or negative mutual dependency principles. We also apply our framework to two real life datasets: Goodreads as a users-rating-books setting and Buzzcity as a users-clicking-advertisements setting. The results show that our method is able to detect suspected spamming users and spammed books in Goodreads and achieve higher precision in identifying fraudulent advertisement publishers than existing approaches.

Keywords-Anomaly Detection; Bipartite Graph; Mutual Dependency; Mutual Reinforcement; Node Anomalies

I. INTRODUCTION

Many real life applications can be modeled as bipartite graphs, including users-rating-products in online marketplaces, users-clicking-webpages on the World Wide Web and users-referring-users in social networks.

In these bipartite graphs, a directed edge carries the "opinion" of a source node towards a target node. For instance, an edge conveys the rating given to a product by a user or the number of times a user has clicked on a webpage.

Moreover, from the perspective of a target node, an edge can be identified as **agreeing** or **disagreeing** by whether the opinion carried by this edge agrees with the majority opinion on the target node. For example in Figure 1, we show a toy example with 5 users (represented by s_1 to s_5) and 3 products (represented by t_1 to t_3) with the edges carrying ratings on a scale of 5. We observe that edge (s_4, t_2) and
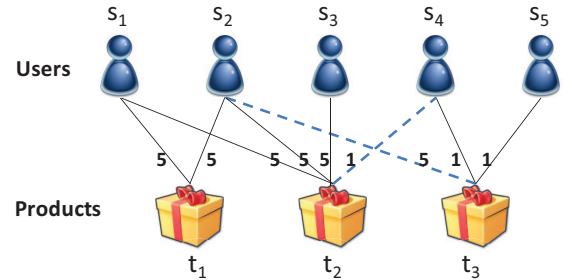

Figure 1. A toy example of 5 users rating 3 products, where the solid lines represent agreeing edges and the dotted lines represent disagreeing edges.

edge (s_2, t_3) do not agree with the majority opinion on target nodes t_2 and t_3 respectively. We denote the two disagreeing edges with dotted lines and the agreeing edges with solid lines.

In general, in a bipartite graph, anomalous nodes are the minority and are inconsistent with the rest of the nodes in the same partite. From this toy example, we first try to manually identify node anomalies that are inconsistent with the rest in terms of opinions. We see from the figure that users s_1, s_2 and s_3 hold the same opinion toward products t_1 and t_2, on the other hand, s_4 and s_5 agree on t_3. However, disagreeing edges (s_4, t_2) and (s_2, t_3) show that $\{s_1, s_2, s_3\}$ and $\{s_4, s_5\}$ hold different opinions on products. Since the majority of the source nodes should be normal, users in $\{s_1, s_2, s_3\}$ are considered to be normal. As a result, products t_1 and t_2 that are agreed upon by the normal users are normal products, whereas the product t_3 agreed by anomalous users are considered to be anomalous. A possible scenario is: $\{s_1, s_2, s_3\}$ are normal users who like products t_1 and t_2. Also s_2 thinks that product t_3 is very good. However, products t_2 and t_3 that are given high ratings by normal users are given very low ratings by s_4 and s_5, who are possibly two spammers hired to demote t_2 and t_3.

In reality, we are not likely to be able to observe a clear split between opinion groups. Rather we need to derive "local" principles for identifying anomalous nodes in both partites. One observation is that we cannot judge a node by its edges alone, instead we should also involve the linked nodes in the other partite. For example, we cannot say s_2 is

anomalous simply because he gives t_3 a minority rating. In fact, it is natural for s_2, a normal node, to give a minority rating to t_3, which is mostly demoted by spammers. In contrast, s_4 giving a minority rating to t_2 should be identified as anomalous. This is because t_2 is a normal product and thus any user who disagrees with the majority who gave high ratings is suspicious. Similarly, we cannot say s_5 is normal simply because he gives t_3 a majority rating, as the fact that t_3 is anomalous makes s_5 giving an agreeing edge to t_3 also anomalous.

Thus, an agreeing edge plays a **positive mutual dependency** role on its source and target. For example, s_1 is more normal if t_1 is more normal and vice versa; similarly, s_5 is more anomalous if t_3 is more anomalous and vice versa. It can also be observed that an disagreeing edge acts as a **negative mutual dependency** channel on its ends. For example, s_2 is more normal if t_3 is more anomalous and vice versa.

We emphasize that both positive and negative mutual dependencies are important. For example, s_5 would not be marked as anonymous without the positive mutual dependency on the anomalous t_3. Further, t_3 is flagged as anomalous only due to the negative mutual dependency on s_2, a normal user.

We therefore arrive at the following integral set of **mutual dependency principles**: (I)**positive mutual dependency** states that a source is more anomalous if it gives agreeing edges to anomalous targets, and more normal if it gives agreeing edges to normal targets. (II)**negative mutual dependency** states that a source is more anomalous if it gives disagreeing edges to normal targets, and more normal if it gives disagreeing edges to anomalous targets. Although the principles are stated in terms of judging source nodes, equivalent principles apply on target nodes.

Previous studies have proposed to model only one of the principles. The mutual reinforcement principle is utilized to rank webpages [1], to identify salient terms and sentences [2], to detect reliable users and contents in social media [3] and to find suspicious reviewers [4]. The negative mutual dependency principle is used to detect biased users and controversial products in evaluation systems [5]. We are the first to propose a generic anomaly detection framework that integrates both sets of mutual dependency principles.

Our contributions are summarized as follows:

- We are the first to propose a generic node anomaly detection framework with mutual dependency principles that capture the complete set of anomalous node behaviors.
- We mathematically formulate our principles and iteratively compute the anomaly scores of source and target nodes. We prove that with our algorithm, the ranking of source and target nodes converges.
- We design an algorithm to generate synthetic bipartite graphs and compare our model with competitors that in-

corporate either positive or negative mutual dependency principles; experiment results show that our model achieves much higher precision.
- We apply our model on two real life datasets: Goodreads and Buzzcity. The results show that we can identify suspected spamming users and spammed books in Goodreads. For the Buzzcity data with labeled ground truth, we also identify fraudulent IP addresses along with the fraudulent advertisement publishers in Buzzcity with higher precision than existing approaches.

The rest of the paper is organized as follows. Section II discusses related work. We formulate our problem and present our model in Section III, and Section IV reports on experiments. We conclude the paper in Section V.

II. RELATED WORK

Our work is related to existing studies on graph anomalies. In [6], a node is considered as anomalous if its neighborhood significantly differs from those of others. The neighborhood of a node is summarized by pairs of neighborhood features, which are assumed to follow power law distribution. The nodes whose neighborhoods deviate from the fitted power law curve are flagged as node anomalies.

[7] detects node anomalies in semantic graphs. The neighborhood of a node is summarized by paths of various length. A node is an anomaly if it carries abnormal semantic paths, which is discovered by the standard distance based anomaly detection techniques.

Anomalous subgraph patterns are studied in [8]. The basic assumption is anomalies induce irregularities in the information content of the graph. They further assume regularities or normal patterns are the best substructures pattern in terms of Minimum Description Length (MDL). Subgraph outliers are the ones experiencing less compressions by best substructure patterns.

In [9], node anomaly in bipartite graph is studied. A source/target node is anomalous if the average similarity of its 1 hop neighbors are low. The similarity of nodes are then computed using random walk.

However, we differ from the aforementioned studies as we detect node anomalies based on the novel mutual dependency principles and we detect anomalies in both partites simultaneously.

Our work is also closely related to studies on mutual dependency of different types of nodes in graphs. [1] proposes the concepts of authority and hub and use the mutual reinforcement principle to rank webpages. The assumption is a good authority is pointed by many good hubs and a good hub points to many good authorities. [10] also uses mutual reinforcement principle to study the veracity of information on the web. They assume a web site is trustworthy if it provides many pieces of true information, and a piece of

information is likely to be true if it is provided by many trustworthy web sites.

Salient terms and sentences are identified using positive mutual dependency principle in [2]. They assume a term has high saliency score if it appears in many sentences with high saliency scores and a sentence has high saliency score if it contains many terms with high saliency scores. [11] extends the model in [2] and proposes mutual reinforcement chain of document, sentence and terms.

In [3], users, questions and answers in the community question answering setting are modeled as three types of nodes. Coupled multiple mutual reinforcing relationships among the three types of nodes are utilized to detect high-quality answers, questions, and users. [4] finds suspicious reviewers in constructed review graph with three types of nodes including reviewers, reviews and stores by positive mutual dependency principle. The trustiness of reviewers, the honesty of reviews, and the reliability of stores are iteratively computed and thus the spam reviews are detected.

[5] uses negative mutual dependency principle to detect bias users and controversial products in evaluation systems. A reviewer is more biased if he deviates more on less controversial objects and an object is more controversial if there is greater deviation by less biased reviewers.

However, our framework is different from existing work as it incorporates both positive and negative mutual dependency principles. Another difference is we prove the ranking of nodes stays unchange after certain number of iterations, whereas existing studies either shows the scores converge to the principle eigenvector of some matrix or only demonstrate the convergence of scores by experiments.

III. ANOMALY DETECTION FRAMEWORK

A. Problem Definition

Given a bipartite graph $G = \langle S \cup T, E, A \rangle$, where $S = \{s_1, ..., s_{|S|}\}$ is a set of nodes in the source partite, $T = \{t_1, ..., t_{|T|}\}$ is a set of nodes in the target partite, $E \subset S \times T$ is a set of directed edges from the source partite to the target partite, and $A = \{a_{ij}\}$ is a set of labels attached to edges, such that

$$a_{ij} = \begin{cases} 0, & \text{if } (s_i, t_j) \text{ is an agreeing edge;} \\ 1, & \text{if } (s_i, t_j) \text{ is a disagreeing edge.} \end{cases}$$

the **anomaly detection problem** is to assign an anomaly score to each node in each partite. The **anomaly score** of a node is a value in $[0, 1]$, with $[0, 0.5)$ being the normal range and $(0.5, 1]$ being the anomalous range. In particular, 0 indicating absolute normality and 1 indicating absolute anomaly.

B. Model Formulation

To facilitate formulation, we summarize our aforementioned principles in Table I.

principles	source	edge	target
positive mutual dependency	normal	agreeing	normal
	anomalous	agreeing	anomalous
negative mutual dependency	normal	disagreeing	anomalous
	anomalous	disagreeing	normal

Table I
MUTUAL DEPENDENCY PRINCIPLES.

We first show how to compute the anomaly score $\mathfrak{s}_i$ of source node s_i from an edge label a_{ij} and the corresponding anomaly score $\mathfrak{t}_j$ of target node t_j. Our principle is that (I)when $a_{ij} = 0$, $\mathfrak{s}_i$ mirrors $\mathfrak{t}_j$, i.e., $\mathfrak{s}_i = \mathfrak{t}_j$; (II)when $a_{ij} = 1$, $\mathfrak{s}_i$ is the opposite of $\mathfrak{t}_j$, i.e., $\mathfrak{s}_i = 1 - \mathfrak{t}_j$. The two conditions together give rise to:

$$\mathfrak{s}_i = \begin{cases} \mathfrak{t}_j, & a_{ij}=0; \\ 1 - \mathfrak{t}_j, & a_{ij}=1. \end{cases}$$

It is easy to see that the above formula is equivalent to : $\mathfrak{s}_i = (1 - 2a_{ij})\mathfrak{t}_j + a_{ij}$.

When it comes to compute the anomaly score $\mathfrak{t}_j$ from $\mathfrak{s}_i$ and a_{ij}, an equivalent formula applies.

$$\mathfrak{t}_j = \begin{cases} \mathfrak{s}_i, & a_{ij}=0; \\ 1 - \mathfrak{s}_i, & a_{ij}=1. \end{cases}$$

We hence have $\mathfrak{t}_j = (1 - 2a_{ij})\mathfrak{s}_i + a_{ij}$, as source and target are symmetric in our principles.

We calculate the anomaly score of a node s_i or t_j as the aggregated anomaly score of all its linked target or source nodes. As we want to account for the impact of all the connected nodes, in the absence of information on the relative importance of various connected nodes, the reasonable option is to give them equal weights. The simple average achieves this, while keeping the score within [0,1]. Furthermore, average allows nice matrix transformation of our formula. Hence we have the following formula:

$$\begin{cases} \mathfrak{s}_i = AVG_{t_j:(s_i,t_j)\in E}(1 - 2a_{ij})\mathfrak{t}_j + a_{ij} \\ \mathfrak{t}_j = AVG_{s_i:(s_i,t_j)\in E}(1 - 2a_{ij})\mathfrak{s}_i + a_{ij} \end{cases} \quad (1)$$

C. Iterative Computation

We can now design the iterative process to compute the anomaly scores of sources and targets by translating formula 1 into a matrix form.

Let $W^S = [w_{ij}^S]$ be a $|S|$ by $|T|$ matrix s.t.

$$w_{ij}^S = \begin{cases} \frac{1}{out\_degree\_of\_s_i}, & \text{if } (s_i, t_j) \in E; \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, Let $W^T = [w_{ji}^T]$ be a $|T|$ by $|S|$ matrix, s.t.,

$$w_{ji}^T = \begin{cases} \frac{1}{in\_degree\_of\_t_j}, & \text{if } (s_i, t_j) \in E; \\ 0, & \text{otherwise.} \end{cases}$$

We then define $X = [x_{ij}]$ as a $|S|$ by $|T|$ matrix s.t.

$$x_{ij} = \begin{cases} a_{ij}w_{ij}^S, & \text{if } (s_i, t_j) \in E; \\ 0, & \text{otherwise.} \end{cases}$$

$Y = [y_{ji}]$ as a $|T|$ by $|S|$ matrix s.t.

$$y_{ji} = \begin{cases} a_{ij}w_{ji}^T, & \text{if } (s_i, t_j) \in E; \\ 0, & \text{otherwise.} \end{cases}$$

Let $\overline{X}$ be a vector with $|S|$ rows s.t. $\overline{x_i} = \sum_{m=1}^{|T|} x_{im}$ and $\overline{Y}$ be a vector with $|T|$ rows s.t. $\overline{y_j} = \sum_{m=1}^{|S|} y_{jm}$.

Let $\mathfrak{S}$ and $\mathfrak{T}$ denote the vectors of anomaly scores of sources and anomaly scores of targets respectively, Formula 1 can be translated to the matrix form as follows:

$$\begin{cases} \mathfrak{S} = (W^S - 2X)\mathfrak{T} + \overline{X} \\ \mathfrak{T} = (W^T - 2Y)\mathfrak{S} + \overline{Y} \end{cases}$$

$\mathfrak{S}$ and $\mathfrak{T}$ are computed iteratively. Let $\mathfrak{S}^k$ and $\mathfrak{T}^k$ denote the vectors of anomaly scores of sources $\mathfrak{s_i}$ and anomaly scores of targets $\mathfrak{t_j}$ in iteration k respectively, then we have the following iterative formula to compute the anomaly scores of source and target nodes.

$$\begin{cases} \mathfrak{S}^k = (W^S - 2X)(W^T - 2Y)\mathfrak{S}^{k-1} + (W^S - 2X)\overline{Y} + \overline{X} \\ \mathfrak{T}^k = (W^T - 2Y)(W^S - 2X)\mathfrak{T}^{k-1} + (W^T - 2Y)\overline{X} + \overline{Y} \end{cases}$$
(2)

D. Convergence

The formula of HITS [1] based on the mutual reinforcement principle can be represented as the eigenvector equation and the iteration computation of hub and authority scores will converge to the principle eigenvector of $A^T A$ and AA^T respectively, where A is the adjacency matrix of the graph. Other existing work including [11] transforms their corresponding matrix to a stochastic and irreducible one by column normalization in order to guarantee convergence. In [5], although their formulation is not in the form of eigenvector equation, they translate their formula into an eigenvector equation by assuming the sum of the anomaly scores of all source or target is 1. To guarantee convergence, they normalize the scores during the iterative computation.

However, in our case, we do not make the assumption about the sum of anomaly scores. Nor can we normalize the corresponding matrix as done in the previous studies. This is because, since any score above 0.5 is considered as anomalous and any score below 0.5 implies normality, normalization may convert a node's score above 0.5 to a score below 0.5, which leads to a "wrong" perception about this node in the iteration process.

In this section, we prove that under certain assumptions, the ranking of nodes in each partite stays unchanged after a certain number of iterations. We only show the proof for source nodes here. The proof for target nodes is similar, as source and target are symmetric in our model.

Let $Q = (W^S - 2X)(W^T - 2Y)$ and $b = (W^S - 2X)\overline{Y} + \overline{X}$. Here, Q is a $|S|$ by $|S|$ matrix, b is a vector with $|S|$

rows and $\mathfrak{S}^k$ is the score vector with $|S|$ rows. Thus, we have $\mathfrak{S}^k = Q\mathfrak{S}^{k-1} + b$.

Lemma 1: For $\mathfrak{S}^k = Q\mathfrak{S}^{k-1} + b$, where $k \geq 1$, if the initial value $\mathfrak{S}^0 = (0.5, 0.5, ..., 0.5)'$, $\mathfrak{S}^k$ always has the solution of $(0.5, 0.5, ..., 0.5)'$, for any $k > 1$.

This lemma is easily proven. According to our formulation, if the anomaly scores of all targets are 0.5, then the anomaly scores of all sources are also 0.5, regardless of the edge label. Therefore, if the initial anomaly score vector for source is $(0.5, 0.5, ..., 0.5)'$, all scores of source nodes stay at 0.5 for any number of iterations.

We then have the following Lemma 2 to transform $\mathfrak{S}^k = Q\mathfrak{S}^{k-1} + b$ to facilitate the convergence study.

Lemma 2: Let e be a $|S|$ dimensional vector with all its elements being 1, i.e. $e = (1, 1, ..., 1)'$. If the largest eigenvalue of Q is smaller than 1, given any initial vector $\mathfrak{S}^0$ such that $\forall\, i, j \leq |S|$ $(i \neq j)$, $\mathfrak{s_i}^0 = \mathfrak{s_j}^0 \neq 0.5$, then $\mathfrak{S}^k = Q\mathfrak{S}^{k-1} + b$ can be represented as $\frac{\mathfrak{S}^k}{\|Q^k e\|_1} = \frac{\theta Q^k e}{\|Q^k e\|_1} + \frac{0.5e}{\|Q^k e\|_1}$ where $-0.5 \leq \theta \leq 0.5$.

Proof: Since $\mathfrak{S}^k = Q\mathfrak{S}^{k-1} + b$, after we substitute $\mathfrak{S}^{k-1}$ by $Q\mathfrak{S}^{k-2} + b$ and then substitute $\mathfrak{S}^{k-2}$ and so on, we arrive at $\mathfrak{S}^k = Q^k\mathfrak{S}^0 + (I + Q + ... + Q^{k-1})b$.

As the largest eigenvalue $\rho(Q) < 1$, we have $\lim_{k \to \infty} Q^k = 0$. $I + Q + ... + Q^k$ is in fact the Neumann Series, which has been shown to converge to $(I - Q)^{-1}$ [12]. Therefore, $\mathfrak{S}^k$ converges to $(I - Q)^{-1}b$ for an arbitrary $\mathfrak{S}^0$.

Let R^k be another anomaly score vector and $R^0 = 0.5e$. According to Lemma 1, we have $R^1 = QR^0 + b = 0.5e$, $R^2 = QR^1 + b = 0.5e$, and in general, $R^k = QR^{k-1} + b = 0.5e$, for every $k \geq 1$.

Since the initial vector $\mathfrak{S}^0$ has equal elements that are not 0.5, we can denote any initial vector as $\mathfrak{S}^0 = (0.5 + \theta)e$. As the initial anomaly score is assumed to be within [0,1], we have $-0.5 \leq \theta \leq 0.5$.

For any θ, $(-0.5 \leq \theta \leq 0.5)$, we have $\mathfrak{S}^1 = Q\mathfrak{S}^0 + b = Q(0.5e + \theta e) + b = QR^0 + b + Q\theta e = R^1 + Q\theta e$. Similarly, we can show that $\mathfrak{S}^2 = R^2 + Q^2\theta e$.

Therefore, we have $\mathfrak{S}^k = \theta Q^k e + R^k = \theta Q^k e + 0.5e$, which can be represented as: $\frac{\mathfrak{S}^k}{\|Q^k e\|_1} = \frac{\theta Q^k e}{\|Q^k e\|_1} + \frac{0.5e}{\|Q^k e\|_1}$. ∎

With Lemma 2, in order to study the convergence of $\mathfrak{S}^k$, we can prove the convergence of $\frac{Q^k e}{\|Q^k e\|_1}$, since other parts of $\frac{\theta Q^k e}{\|Q^k e\|_1} + \frac{0.5e}{\|Q^k e\|_1}$ are of known value.

Lemma 3: $\frac{Q^k e}{\|Q^k e\|_1}$ converges when $k \to \infty$.

Proof: Let $\sigma(Q) = \{\lambda_1, \lambda_2, ..., \lambda_x\}$ denote x number of distinct eigenvalues of Q. According to [12], there exists a nonsingular matrix P s.t.

$$J = P^{-1}QP = \begin{pmatrix} J(\lambda_1) & 0 & \cdots & 0 \\ 0 & J(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J(\lambda_x) \end{pmatrix},$$

J is called the Jordan form of Q, and each of the $J(\lambda_i)$ takes

the form of $\begin{pmatrix} J_1(\lambda_i) & 0 & \cdots & 0 \\ 0 & J_2(\lambda_i) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_y(\lambda_i) \end{pmatrix}$, where y

can be calculated as in [12],

$$J_\star(\lambda_i) = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}_{m \times m}$$, where m can be

calculated as in [12]. Therefore, we have $Q = PJP^{-1}$, and $Q^k = PJ^kP^{-1}$. Since J is block diagonal, we have

$$J^k = \begin{pmatrix} J(\lambda_1)^k & 0 & \cdots & 0 \\ 0 & J(\lambda_2)^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J(\lambda_x)^k \end{pmatrix}.$$

Since $J(\lambda_i)$ is also block diagonal, we are interested to know the form of $J_\star(\lambda_i)^k$ with $m \times m$, which can be shown as

$$J_\star(\lambda_i)^k = \begin{pmatrix} \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} & \cdots & \binom{k}{m-1}\lambda_i^{k-m+1} \\ & \lambda_i^k & \ddots & \vdots \\ & & \ddots & \binom{k}{1}\lambda_i^{k-1} \\ & & & \lambda_i^k \end{pmatrix}$$

Let D be the set of distinct terms that appear in at least one element of $Q^k e$. Each term in D is of form $\binom{k}{m-1}\lambda^{k-m+1}$, where $m \geq 1$ and $\lambda \in \sigma(Q)$.

Therefore, $Q^k e = \begin{pmatrix} \vdots \\ \sum_j c_{i,j} d_{i,j} \\ \vdots \end{pmatrix}$. Here i represents the

i-th element of $Q^k e$, j represents the j-th term of an element, and $c_{i,j}$ is a non-zero real value.

It can be easily proven as follows that there exists a $d^* \in D$, such that $\forall d \in D - \{d^*\}$, $\lim_{k \to \infty} \frac{|d|}{|d^*|} = 0$. For any given two terms in D, we have

(I) if the terms contain the same λ, then $\binom{k}{m_1}/\binom{k}{m_2} \to 0$, when $k \to \infty$, $m_1 < m_2$.

(II) if the terms of an element contain different λ_1 and λ_2, then $\binom{k}{m}(\frac{\lambda_1}{\lambda_2})^k \to 0$, when $k \to \infty$, $\lambda_1 < \lambda_2$.

Therefore, there exists a $d^* \in D$ such that $\forall d \in D - \{d^*\}$, $\lim_{k \to \infty} \frac{|d|}{|d^*|} = 0$. As $\frac{Q^k e}{\|Q^k e\|_1} = \frac{Q^k e}{\sum_i |\sum_j c_{i,j} d_{i,j}|} = \frac{Q^k e/|d^*|}{\sum_i |\sum_j c_{i,j} d_{i,j}/d^*|}$, we have, $\lim_{k \to \infty} \frac{Q^k e}{\|Q^k e\|_1} = \lim_{k \to \infty} \frac{Q^k e/|d^*|}{\sum_i |\sum_j c_{i,j} d_{i,j}/d^*|}$.

For the denominator, for the terms $d_{i,j} = d^*$, the limit of $d_{i,j}/d^*$ is 1, whereas for the terms $d_{i,j} \neq d^*$, $d_{i,j}/d^*$ is 0. Hence, the limit of the denominator is the sum of the $c_{i,j}$ where the corresponding $d_{i,j} = d^*$. On the other hand, the limit of the numerator is a vector of real values where only the elements whose terms contain d^* is non-

zero value. Hence, the whole fraction converges. Therefore, $\frac{Q^k e}{\|Q^k e\|_1}$ converges when $k \to \infty$. ∎

With Lemma 3, we have the following Theorem 1 regarding the convergence of anomaly scores $\mathfrak{S}^k$:

Theorem 1: For $\mathfrak{S}^k = Q\mathfrak{S}^{k-1} + b$, where $k \geq 1$, if the largest eigenvalue of Q is smaller than 1, given any initial vector $\mathfrak{S}^0$ such that $\forall i, j \leq |S|$ ($i \neq j$), $\mathfrak{s}_i^0 = \mathfrak{s}_j^0 \neq 0.5$, then $\mathfrak{S}^k$ converges when $k \to \infty$.

With Lemma 2 and Lemma 3, this theorem is obvious.

We now study the convergence of rankings of nodes according to their anomaly scores $\mathfrak{S}^k$. We have the following:

Theorem 2: For $\mathfrak{S}^k = Q\mathfrak{S}^{k-1} + b$, where $k \geq 1$, if the largest eigenvalue of Q is smaller than 1, given any initial vector $\mathfrak{S}^0$ such that $\forall i, j \leq |S|$ ($i \neq j$), $\mathfrak{s}_i^0 = \mathfrak{s}_j^0 \neq 0.5$, then $\exists$ an integer $K > 0$ such that the ranking of elements in $\mathfrak{S}^k$ will stay unchanged $\forall k > K$.

Proof: Since Q's largest eigenvalue is smaller than 1, and the initial vector $\mathfrak{S}^0$ contains identical values not equal to 0.5, according to Lemma 2, we have $\frac{\mathfrak{S}^k}{\|Q^k e\|_1} = \frac{\theta Q^k e}{\|Q^k e\|_1} + \frac{0.5e}{\|Q^k e\|_1}$. Let $\frac{Q^k e}{\|Q^k e\|_1}(i)$ denote the i-th element of vector $\frac{Q^k e}{\|Q^k e\|_1}$. Let $\Delta = min_{i,j}(\frac{Q^k e}{\|Q^k e\|_1}(i) - \frac{Q^k e}{\|Q^k e\|_1}(j))$, then for any i, j, we have $|\frac{Q^k e}{\|Q^k e\|_1}(i) - \frac{Q^k e}{\|Q^k e\|_1}(j)| \geq \Delta$.

According to Lemma 3, $\frac{Q^k e}{\|Q^k e\|_1}$ converges. Hence, for any real number $\varepsilon > 0$, $\exists$ an integer $K > 0$, s.t., $\forall k > K$, $max_i(|\frac{Q^{k+1} e}{\|Q^{k+1} e\|_1}(i) - \frac{Q^k e}{\|Q^k e\|_1}(i)|) < \varepsilon$. Therefore, if ε is smaller than $\Delta/2$, the above still holds.

Suppose $\frac{Q^k e}{\|Q^k e\|_1}(i) > \frac{Q^k e}{\|Q^k e\|_1}(j)$, for any $\varepsilon < \Delta/2$, we have $\frac{Q^{k+1} e}{\|Q^{k+1} e\|_1}(i) > \frac{Q^k e}{\|Q^k e\|_1}(i) - \varepsilon$, and $\frac{Q^{k+1} e}{\|Q^{k+1} e\|_1}(j) < \frac{Q^k e}{\|Q^k e\|_1}(j) + \varepsilon$. Since $|\frac{Q^{k+1} e}{\|Q^{k+1} e\|_1}(i) - \frac{Q^k e}{\|Q^k e\|_1}(i)| < \varepsilon$ and $|\frac{Q^{k+1} e}{\|Q^{k+1} e\|_1}(j) - \frac{Q^k e}{\|Q^k e\|_1}(j)| < \varepsilon$, we always have $\frac{Q^{k+1} e}{\|Q^{k+1} e\|_1}(i) > \frac{Q^{k+1} e}{\|Q^{k+1} e\|_1}(j)$.

Therefore, the relative order of $\frac{Q^k e}{\|Q^k e\|_1}(i)$ and $\frac{Q^k e}{\|Q^k e\|_1}(j)$ stays unchanged for all $k > K$. Therefore, we have proven that there exists an integer $K > 0$, s.t. the ranking of all elements in $\mathfrak{S}^k$ never changes $\forall k > K$. ∎

With the proven Theorem 2, we know that with certain assumption about Q and as long as we set the identical initial value for all source nodes, the ranking of source nodes will stay the same after a certain number of iterations. Now the question is, will different runs with initial vectors of different identical values converge to the same ranking?

The following theorem shows that (I)if two different runs involve different identical initial values that are both smaller 0.5, the final rankings are the same; (II)if two different runs involve different identical initial values that are both greater than 0.5, the final rankings are the same; (III)if one run involves initial values smaller than 0.5 and another run involves initial values greater than 0.5, the final rankings are the opposite.

Theorem 3: For $\mathfrak{S}^k = Q\mathfrak{S}^{k-1} + b$, where $k \geq 1$, if the largest eigenvalue of Q is smaller than 1, we have:
(I) given one initial vector $\mathfrak{S}^0$, such that $\forall\ i,j \leq |S|\ (i \neq j)$, $\mathfrak{s_i}^0 = \mathfrak{s_j}^0 < 0.5$, and another initial vector $\mathfrak{S}^{*0}$, such that $\forall\ i,j \leq |S|\ (i \neq j)$, $\mathfrak{s_i}^0 = \mathfrak{s_j}^0 < 0.5$, then there exist an integer $K > 0$, such that the two rankings $\mathfrak{S}^k$ and $\mathfrak{S}^{*k}$ are identical $\forall k > K$.
(II) given one initial vector $\mathfrak{S}^0$, such that $\forall\ i,j \leq |S|\ (i \neq j)$, $\mathfrak{s_i}^0 = \mathfrak{s_j}^0 > 0.5$, and another initial vector $\mathfrak{S}^{*0}$, such that $\forall\ i,j \leq |S|\ (i \neq j)$, $\mathfrak{s_i}^0 = \mathfrak{s_j}^0 > 0.5$, then there exist an integer $K > 0$, such that the two rankings $\mathfrak{S}^k$ and $\mathfrak{S}^{*k}$ are identical $\forall k > K$.
(III) given one initial vector $\mathfrak{S}^0$, such that $\forall\ i,j \leq |S|$ $(i \neq j)$, $\mathfrak{s_i}^0 = \mathfrak{s_j}^0 < 0.5$, and another initial vector $\mathfrak{S}^{*0}$, such that $\forall\ i,j \leq |S|\ (i \neq j)$, $\mathfrak{s_i}^0 = \mathfrak{s_j}^0 > 0.5$, then there exist an integer $K > 0$, such that the two rankings $\mathfrak{S}^k$ and $\mathfrak{S}^{*k}$ are exactly the opposite $\forall k > K$.

Proof: Since Q's largest eigenvalue is smaller than 1, and the initial vector $\mathfrak{S}^0$ contains identical values not equal to 0.5, according to Lemma 2, we have $\frac{\mathfrak{S}^k}{\|Q^k e\|_1} = \frac{\theta Q^k e}{\|Q^k e\|_1} + \frac{0.5e}{\|Q^k e\|_1}$, where $-0.5 \leq \theta \leq 0.5$.

Since the initial vector $\mathfrak{S}^0$ has equal elements not equal to 0.5, we have denoted in Lemma 2 any initial vector as $\mathfrak{S}^0 = (0.5+\theta)e$. When $\theta > 0$, the initial vector has elements larger than 0.5, whereas when $\theta < 0$, the initial vector has elements smaller than 0.5.

Therefore, (I) actually states $\mathfrak{S}^0 = (0.5 + \theta)e$, $\mathfrak{S}^{*0} = (0.5 + \theta^*)e$ with $\theta < 0$ and $\theta^* < 0$. (II) involves $\theta > 0$ and $\theta^* > 0$ and (III) suggests $\theta < 0$ and $\theta^* > 0$.

Thus, we have $\mathfrak{S}^k = \theta Q^k e + 0.5e$ and $\mathfrak{S}^{*k} = \theta^* Q^k e + 0.5e$. It is easy to see that if (I)$\theta < 0$ and $\theta^* < 0$ or (II)$\theta > 0$ and $\theta^* > 0$, $\mathfrak{S}^k$ will only differ from $\mathfrak{S}^{*k}$ by some scale. The rankings of all elements will be the same. When (III)$\theta < 0$ and $\theta^* > 0$, any two elements in $\mathfrak{S}^k$ will have a reverse ranking in $\mathfrak{S}^{*k}$. Thus $\mathfrak{S}^k$ is a reverse ranking of $\mathfrak{S}^{*k}$.

According to Theorem 2, suppose K and K^* are the two integers where the rankings of $\mathfrak{S}^k$ and $\mathfrak{S}^{*k}$ stay unchanged, we know $\forall k > \max(K, K^*)$, both rankings will stay unchanged. Therefore, we have proven (I), (II) and (III). ∎

Theorem 3 suggests that if we set the initial value to be smaller than 0.5, implying originally all nodes are normal, we will get the ranking that is exactly the opposite of that we get if we set the initial value to be larger than 0.5, implying originally all nodes are anomalous. Thus, the initial values act as the "prior view" towards all nodes. Since we assume the normal nodes are the majority, we therefore should always set the initial value to be smaller than 0.5.

E. Iterative Algorithm

With these Theorems we now design our iterative algorithm, Algorithm 1 to compute the ranking of source nodes. The ranking of target nodes can be computed similarly.

Algorithm 1 *Iterative algorithm to compute ranking of source nodes.*

Input: bipartite graph $G = \langle S \cup T, E, A \rangle$, initial value x, number of iterations ranking stays unchanged y.
Output: rankings of S, $\widehat{S}$.
1: Set $\mathfrak{S}^0 = (x, x, ..., x)'$.
2: Let $\widehat{S^*}$ be the ranking of S according to $\mathfrak{S}^0$.
3: $c=0$; $k=1$.
4: **while** $c \leq y$ **do**
5:      Compute anomaly score for source nodes as in Formula 2,
     $\mathfrak{S}^k = (W^S - 2X)(W^T - 2Y)\mathfrak{S}^{k-1} + (W^S - 2X)\overline{Y} + \overline{X}$.
6:      $\widehat{S}$ keeps the ranking of S according to $\mathfrak{S}^k$.
7:      **if** the distance between $\widehat{S}$ and $\widehat{S^*}$ is 0 **then**
8:          c++.
9:      **else**
10:          $c = 0$.
11:      $\widehat{S^*} = \widehat{S}$.
12:      k++.
13: **return** $\widehat{S}$

We set the initial score vector of both source and target nodes as $(0.1, 0.1, ..., 0.1)'$. We measure the rankings of the two consecutive iterations by Kullback-Leibler divergence [13]. In practice, the actual number of iterations needed is small, which is studied in the experiments section.

IV. EXPERIMENTS

In this section, we evaluate our node anomaly detection framework on both synthetic data and real life data. We generate synthetic bipartite graphs with properties that are necessary for testing different detection models. We show the precisions of our method as well as other existing methods on these synthetic graphs. Experiments on two real life data, namely Goodreads and Buzzcity showcase the ability to identify suspicious spamming users and spammed books in Goodreads and to identify fraudulent users and advertisement publishers in Buzzcity with higher precision than existing approaches, both unsupervised and supervised.

A. On Synthetic Data

1) Synthetic Data Generation Algorithm: We first set aside 4 sets of nodes as anomalous source nodes S^A, anomalous target nodes T^A, normal source nodes S^N and normal target nodes T^N and then generate edges among them. In order for an injected anomalous node to be anomalous, it has to possess at least one kind of anomalous characteristics: giving/receiving disagreeing edges to/by normal nodes or giving/receiving agreeing edges to/by anomalous nodes. Similarly, for an injected normal node to be normal, it has to have at least one kind of normal characteristics: giving/receiving disagreeing edges to/by anomalous nodes or giving/receiving agreeing edges to/by normal nodes.

We generate edges in the following sequence: from S^A to T^A, from S^A to T^N, from S^N to T^A, and then from S^N to T^N. We introduce parameters in the generation algorithm so that it generates graphs of different properties (e.g., the ratio of disagreeing edges to agreeing edges), which are necessary
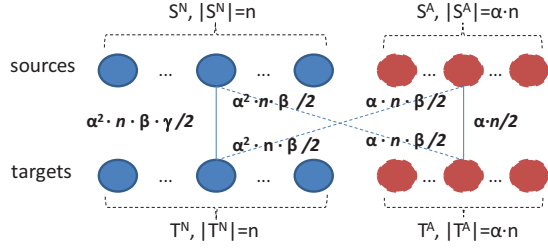
Figure 2. An illustration of the synthetic data generation and its properties. In the graph, there are four sets of nodes, anomalous source nodes S^A, anomalous target nodes T^A, normal source nodes S^N and normal target nodes T^N. The values on the edges represent the expectation of the number of edges from one node set to another.

to test our method as well as others'. We also make sure that in the algorithm generate these properties for source and target nodes simultaneously so that later steps would not mess up the properties generated by the previous ones. Parameter n controls the size of S^N and T^N, α is the ratio of $|S^A|$ to $|S^N|$ or the ratio of $|T^A|$ to $|T^N|$, β is the ratio of the expected number of disagreeing edges to the expected number of agreeing edges for any anomalous node, and $1/\gamma$ is the ratio of the expected number of disagreeing edges to the expected number of agreeing edges for any normal node. The detailed algorithm is shown in Algorithm 2 followed by the analysis and the explanation of the graph properties controlled by these parameters.

Algorithm 2 *Generate synthetic bipartite graph*

Input: parameters: n, α, β, γ
Output: bipartite graph, G
1: Generate node sets for G: S^N, S^A, T^N, T^A such that $|S^N| = |T^N| = n$, $|S^A| = |T^A| = \alpha \cdot n$, $(0 < \alpha < 1)$
2: **for** each node s in S^A **do**
3:   Randomly select a target node set $T^* \subset T^A$, where $|T^*|$ is randomly drawn from $[0, \alpha \cdot n]$, $(\alpha \cdot n \le |T^A| = \alpha \cdot n)$.
4:   Generate agreeing edges from s to each node in T^*.
5:   Randomly select a target node set $T^* \subset T^N$, where $|T^*|$ is randomly drawn from $[0, \alpha \cdot n \cdot \beta]$, $(\alpha \cdot n \cdot \beta \le |T^N| = n)$.
6:   Generate disagreeing edges from s to each node in T^*.
7: **for** each node t in T^A **do**
8:   Randomly select a source node set $S^* \subset S^N$, where $|S^*|$ is randomly drawn from $[0, \alpha \cdot n \cdot \beta]$, $(\alpha \cdot n \cdot \beta \le |S^N| = n)$.
9:   Generate disagreeing edges from each node in $|S^*|$ to t.
10: **for** each node s in S^N **do**
11:   Randomly select a target node set $T^* \subset T^N$, where $|T^*|$ is randomly drawn from $[0, \alpha^2 \cdot n \cdot \beta \cdot \gamma]$, $(\alpha^2 \cdot n \cdot \beta \cdot \gamma \le |T^N| = n)$.
12:   Generate agreeing edges from s to each node in $|T^*|$.
13: **return** G

Figure 2 illustrates the properties of our generated graphs. Specifically, after step 6, it can be easily shown that for any anomalous source, the expectation of the number of agreeing edges is $\alpha \cdot n/2$; and the expectation of the number of disagreeing edges is $\alpha \cdot n \cdot \beta/2$.

We can also compute for any node in T^N, the expected number of edges linking from S^A as: $\frac{1}{\alpha \cdot n \cdot \beta + 1} \cdot \sum_{i=0}^{\alpha \cdot n \cdot \beta} \frac{i}{n} \cdot \alpha \cdot n = \alpha^2 \cdot n \cdot \beta/2$.

Similarly, since $|S^A| = |T^A|$, for any node in T^A, the expected number of edges linking from S^A is exactly $\alpha \cdot n/2$.

After step 9, for any anomalous target in T^A, the expectation of the number of edges from S^N is $\alpha \cdot n \cdot \beta/2$. Similarly, for any node in S^N, the expected number of edges to T^A is $\alpha^2 \cdot n \cdot \beta/2$.

After step 12, we can compute that for any node in S^N, the expectation of the number of edges to T^N is $\alpha^2 \cdot n \cdot \beta \cdot \gamma/2$. Since $|S^N| = |T^N|$, for any node in T^N, the expectation of the number of edges from S^N is also $\alpha^2 \cdot n \cdot \beta \cdot \gamma/2$.

Now, for any node in S^A or in T^A, the ratio of the expected number of disagreeing edges to the expected number of agreeing edges is $\frac{\alpha \cdot n \cdot \beta/2}{\alpha \cdot n/2} = \beta$. On the other hand, for any node in S^N or in T^N, the ratio of the expected number of disagreeing edges to the expected number of agreeing edges is $\frac{\alpha^2 \cdot n \cdot \beta/2}{\alpha^2 \cdot n \cdot \beta \cdot \gamma/2} = 1/\gamma$. The controlling of these ratios is therefore achievable by varying parameters.

In bipartite graphs, each edge carries the opinion of source node to target node. It is natural to expect that the opinions generated by the normal nodes prevail the opinions generated by the anomalous nodes, we should be able to set the parameters so as to control the opinions from anomalous nodes always being minority. Specifically, since disagreeing edges are always between anomalous and normal nodes, the opinions from normal and anomalous nodes carried by disagreeing edges are always equal. Thus, we want the opinions in the form of expected number of agreeing edges among normal nodes to be larger than those among anomalous nodes.

The expected number of agreeing edges among normal nodes is $(|S^N| + |T^N|) \cdot \alpha^2 \cdot n \cdot \beta \cdot \gamma/2 = \alpha^2 \cdot n^2 \cdot \beta \cdot \gamma$. The expected number of agreeing edges among anomalous nodes is $(|S^A| + |T^A|) \cdot \alpha \cdot n/2 = \alpha^2 \cdot n^2$. According to the anomaly being minority assumption, we have $\alpha^2 \cdot n^2 \cdot \beta \cdot \gamma > \alpha^2 \cdot n^2$, which leads to $\beta > 1/\gamma$. Thus, we set β to be larger than $1/\gamma$ to guarantee the graphs are properly generated.

Other parameter constraints can be derived from the generation algorithm. They are $\alpha \cdot n \cdot \beta \le 1$ derived from step 5 and step 8 and $\alpha^2 \cdot \beta \cdot \gamma \le 1$ derived from step 11.

2) Results: We compare our model denoted as IMD with the models that are based on only one of the mutual dependency principles. We denote the one with positive mutual dependency principle as PMD implemented as in [2] and the one with negative mutual dependency principle as NMD implemented as in [5]. We also incorporate the random guess method denoted as RG. We vary the parameters and test the precisions of all methods.

We first test with a fixed α and β, how $1/\gamma$ affects the results. Note that both β and $1/\gamma$ can be set as any real number. If they are smaller than one, agreeing edges dominate for each node. If they are larger than one, disagreeing edges dominate. We set $|S^N| = |T^N| = n$ as 200 and set $\alpha = 0.5$. β can be derived as smaller than or equal to 2. We set $\beta=2$ and set $1/\gamma=\{1.9, 1.5, 1, 0.5\}$ to test on graphs where agree-
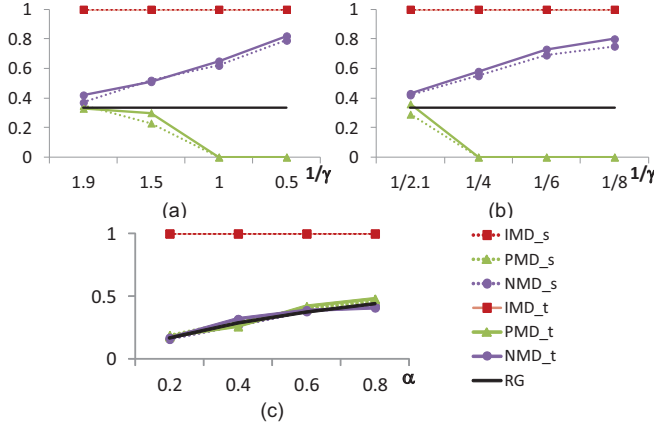
Figure 3. Results on synthetic data. Y-axis shows the average precision@K and X-axis shows different $1/\gamma$ values. (a) with α=0.5, β=2, varying $1/\gamma$; (b) with α=0.5, β=0.5 and varying $1/\gamma$; (c) with $\beta = 0.5$, $1/\gamma = 0.49$ and varying α.

Figure 4. Results on synthetic data with varying noise level π.

ing edges dominate for anomalous nodes. We also try β=0.5 and vary $1/\gamma$={1/2.1, 1/4, 1/6, 1/8} to test on graphs where agreeing edges dominate for anomalous nodes. The values for $1/\gamma$ satisfy the aforementioned parameter constraints. For each parameter setting, we generate 3 bipartite graphs and measure each method by precision@K, where K is the number of true anomalous source/target nodes. Our results with β=2 is shown in Figure 3(a) and the results with β=0.5 in Figure 3(b). Y-axis shows the average precision@K and X-axis shows different $1/\gamma$ values. The curve with IMD_s is regarding the performance of method IMD on source nodes. Similarly, IMD_t is for target nodes.

We can see from the figures that our method IMD gets perfect precision over all settings, much better than the competing ones. When β is fixed and $1/\gamma$ gets smaller, anomalous nodes are getting more characterized by disagreeing edges and less characterized by agreeing edges compared to normal ones. As a result, it becomes easier for the model NMD that propagates anomaly scores through disagreeing edges to identify anomalous nodes. At the same time it becomes harder for the model PMD that propagates anomaly scores through agreeing edges to identify anomalous nodes.

As IMD takes into consideration both mutual dependency principles and propagate anomaly scores on both types of edges, we are able to achieve the perfect precision. This shows that both the positive and negative mutual dependency relationships are necessary for anomaly detection.

To test whether the performance will be affected by the number of anomalous nodes α, we set $\beta = 0.5$ and $1/\gamma = 0.49$. The results are shown in Figure 3(c), which suggests that, as α changes, the performance of NMD and PMD are only as good as the random guess. The explanation is when $1/\gamma$ is only slightly smaller than β, the normal nodes and the anomalous nodes are giving out almost the same ratio of
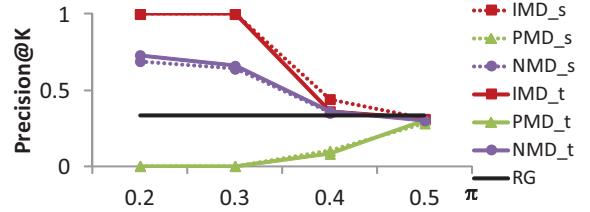
number of agreeing edges to disagreeing edges. This makes the anomalous nodes hard to discriminate from the normal ones for NMD and PMD.

3) Results with Noisy Ground Truth: We have shown the results on synthetic graphs generated under the assumption that normal nodes never behave anomalously and anomalous nodes never show any normal behaviors. This assumption makes sure that the anomalous and normal nodes are truly the ground truth. Here we also experiment on synthetic bipartite graphs with noisy ground truth. In other words, anomalous nodes may possess some normal behaviors: giving/receiving disagreeing edges to/by anomalous nodes or giving/receiving agreeing edges to/by normal nodes; normal nodes may show anomalous behaviors: giving/receiving disagreeing edges to/by normal nodes or giving/receiving agreeing edges to/by anomalous nodes.

Parameter π is used to control the probability of a node having the behavior of its opposing role. Specifically, we modify the Algorithm 2 such that when agreeing edges are generated in step 4 and step 12, each edge has a probability of π of being a disagreeing edge. Similarly, when disagreeing edges are generated in step 6 and step 9, each edge has a probability of π of being an agreeing edge. Note that the larger the π is, the less likely the anomalous and normal nodes are the real ground truth.

We vary π and set other parameters as $n = 200$, $\alpha = 0.5$, $\beta = 0.5$, $1/\gamma = 0.5$, as the competing method performs better on this setting. The results are shown in Figure 4. As we can see that, even in the presence of noise, IMD is still the best. Even when the noisy level is 0.3, IMD resists all noisy information. When $\pi = 0.5$, all methods are as good as the random guess. This is because, when anomalous nodes manifest the same amount of anomalous behavior as normal behaviors, anomalous nodes are no longer anomalous.

4) Convergence Speed: Here we study during the iterative computation, how fast the rankings converge. We measure the distance between the rankings of two consecutive iterations by Kullback-Leibler divergence [13]. If two rankings are the same, the KL divergence distance is 0. We set $\alpha = 0.5$, $\beta = 0.5$, $1/\gamma = 0.5$ and $\pi = 0.1$. We vary the size of n from small to large, such that the total number of nodes in the generated graphs are from 100 to 10,000. The KL divergence distance v.s. the number of iterations curves
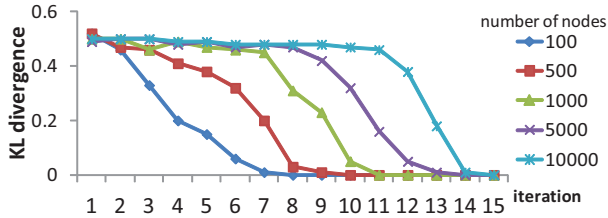
Figure 5. Results on convergence. Different curves corresponding to different graph sizes (number of nodes) show how the KL divergence between the node rankings of two consecutive iterations changes as the number of iterations increases.

for sources are shown in Figure 5. The curves for targets are similar and are omitted due to space constraint. Note that the figure shows the number of iterations until the distance becomes 0. In practice, our algorithm runs 10 more iterations to guarantee the distance does stay as 0 afterwards. We can see from the plot that our node ranking stays unchanged after only a small number of iterations.

B. On Real-life Data

1) Goodreads Data: Goodreads[1] is the largest website for people to write book reviews, rate books, recommend books to friends and socialize online with other readers or writers. According to its website, it has more than 9,000,000 members who have added more than 320,000,000 books to their shelves. To facilitate exploration, Goodreads create "listopia" for users to quick find interesting books and maybe vote for their favorites. A book can be rated by any user on a scale of 5.

We utilize the API provided by Goodreads to crawl our data on May 20, 2012. Since the number of books and users are too large, we start from a popular book list in Goodreads' listopia, called "Dealbreakers: If You Like This Book, We Won't Get Along". We crawl from this list that books that have less than 1000 user ratings. We then crawl all the users who have rated these books and have rated less than 1000 books. We further go another hop to get all the books that are rated by the previously crawled users and have less than 1000 user ratings. The bipartite graph is thus generated with users as the source partite, books as the target partite and the edges suggests users rating books.

We only crawl the books and users with less than 1000 neighbors so that we would not run into some really popular books and end up with too large a dataset. The other reason is that the anomalous books and users with fewer ratings are easier to examine and make sense of.

As our model takes in edge labels which are either agreeing or disagreeing, we thus map the rating carried by an edge to a label as follows. If this rating is a minority rating of all ratings given to the same book, then this edge

is given a disagreeing label; otherwise, this edge is given a agreeing label. The majority ratings are the ones within 2 standard deviation from the mean of all ratings given to a target.

We iteratively filter away target nodes with less than 10 edges and source nodes with less than 2 edges, as books rated by less than 10 users are not drawing enough attention from the audience and less likely to be spammed. Similarly users only rate one book are not influential enough, even if they are spammers. In the end, we have a bipartite graph with 7982 source nodes, 9169 target nodes and 163621 edges.

We apply our anomaly detection method on this bipartite graph. The top-1 anomalous book returned is "Justin Bieber: My World", published in August 1st 2010, and written by Justin Bieber. It received 100 ratings, with 48 of score 5, 4 of score 4, 4 of score 3, 3 of score 2 and 41 of score 1. Judging by this rating distribution, we know the opinions on this book are rather divided. Moreover, it is very hard to tell whether the users giving out 1 or the users giving 5 are spamming or anomalous simply by rating distribution. However, our approach is able to identify out of 99 users, 25 of them are anomalous users (i.e., users whose anomaly scores are greater than 0.5). Interestingly, these anomalous users are all giving low ratings (1 or 2) to this book and none of them gives textual comments to this book. Our model considers these 25 users anomalous as when they rate other books, they tend to disagree with normal users and agree with other anomalous users. We find that some of the 25 users including Angela and Kimiko, who rated around twenty books and mostly gave low ratings are not quite far away from the average rating. Based on our results, we may think that the book "Justin Bieber: My World" is unfairly demoted by some users.

Top-2nd is also a book by Justin Bieber, called "I ♡ Justin Bieber". It received 23 ratings, with 22 of them being 5. Unlike the top-1 book, this book is anomalous, as there are 9 identified anomalous users giving out score 5 to it. Almost all 9 users seem to be the fans of Justin Bieber. Their comments are about loving the person, not the book. One of them even has the user name as "JustinBieberLover". We therefore conclude that this book is somewhat spammed by his fans to promote the book.

Other top 10 anomalous books include 3 other Justin Bieber's books, "Kardashian Konfidential" by Kourtney Kardashian and "Birth Control Is Sinful in the Christian Marriages and Also Robbing God of Priesthood Children!!" by Eliyzabeth Yanne Strong-Anderson. All these books are identified by our model as being rated by some suspicious users.

2) Buzzcity data: BuzzCity[2] is a global mobile advertising platform, where publishers host on their own websites the advertisement of advertisers. If anyone clicks on the

---

[1] http://www.goodreads.com/

[2] http://www.buzzcity.com/

advertisements, the publishers and the BuzzCity platform get money paid by advertisers. It is suspected that some fraudulent publishers would hire spammers to click on the advertisement hosted on their websites to gain money. To maintain a healthy ecosystem, BuzzCity has good intention to identify these spammers.

Buzzcity provides around 10 million click logs during a three day's period from Jan 26, 2012 to Jan 28, 2012. The dataset contains the encoded IP address of each click on each publisher. As a publisher may have multiple websites, the click data is already aggregated by publishers. Buzzcity has asked its own employees to label the publishers as "OK", "Observation" (meaning not sure) or "Fraud". This dataset will be made public for a fraud detection competition.

We apply our model to detect the fraudulent clickers (i.e., IP addresses) and publishers with clickers as one partite, publishers as the other partite and edges suggest users click on publishers. Since each edge carries the number of clicks from a clicker to a publisher, we map the number of clicks to agreeing and disagree edges as we do for the Goodreads data. We iteratively filter out nodes with less than 2 edges, as they are unlikely to be fraudulent. As a result, we have 132540 source nodes, 1428 target nodes, among which 1264 are OK, 77 are Observation and 87 are Fraud cases. Thus a random guess may achieve a precision of around 0.06.

The completing methods of our IMD are PMD with positive mutual dependency, NMD with negative mutual dependency, the distance-based anomaly detection approach DIST and a supervised approach, decision tree DT. For DIST, the distance between two publishers is defined as the KL divergence distance between the corresponding two click distributions of publishers. A click distribution of a publisher is the normalized histogram on the number of clicks from all clickers of this publisher. The distance-based anomaly score is computed as in [14]. As for DT, for each publisher, we define 7 features including total number of unique clickers, total number of clicks, the ratio of number of clicks to the total number of unique clickers, as well as mean, median, standard deviation and skewness of the click distribution. We use C4.5 implemented by WEKA [15] with 10 fold cross-validation and output the prediction score of each publisher.

Since each method can output a ranking of the publishers according to its computed scores, we thus measure the performance of the top ranked publishers of each method by Precision@K. We vary K from small to as large as 87, the number of labeled fraudulent publishers. The results are shown in Figure 6. As we can see from the plot that our method IMD performs the best among all approaches.

V. CONCLUSIONS

We proposed a generic anomaly detection framework on bipartite graphs, based on the integral set of mutual dependency principles. We are the first to unify the positive and negative mutual dependency principles and design iterative
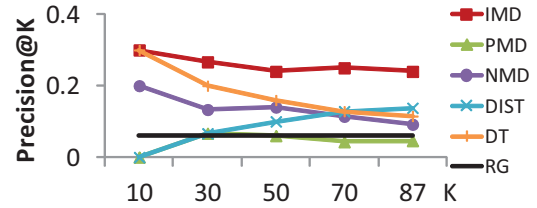

Figure 6.   Precision@K curves on Buzzcity dataset.

algorithm with guarantee that the ranking of sources or targets will stay unchanged for a certain number of iterations. We tested our framework on both the synthetic data and two real life datasets, namely Goodreads and Buzzcity. The results in these datasets show that our model outperforms the models with either positive or negative mutual dependency principles, which demonstrates the necessity of incorporating both principles for anomaly detection tasks. Moreover, we successfully identified suspicious users and books in Goodreads and achieved higher precision in detecting fraudulent publishers in Buzzcity than existing approaches.

REFERENCES

[1] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46(5), 1999.

[2] H. Zha, "Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering," in *SIGIR Conf.*, 2002.

[3] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha, "Learning to recognize reliable users and content in social media with coupled mutual reinforcement," in *WWW Conf.*, 2009.

[4] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection," in *ICDM Conf.*, 2011.

[5] H. W. Lauw, E.-P. Lim, and K. Wang, "Bias and controversy in evaluation systems," *IEEE Trans. on Knowl. and Data Eng.*, vol. 20(11), 2008.

[6] L. Akoglu, M. Mcglohon, and C. Faloutsos., "Oddball: Spotting anomalies in weighted graphs," in *PAKDD Conf.*, 2010.

[7] S.-d. Lin and H. Chalupsky, "Discovering and explaining abnormal nodes in semantic graphs," *IEEE Trans. on Knowl. and Data Eng.*, vol. 20(8), 2008.

[8] C. C. Noble and D. J. Cook, "Graph-based anomaly detection," in *KDD Conf.*, 2003.

[9] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, "Neighborhood formation and anomaly detection in bipartite graphs," in *ICDM Conf.*, 2005.

[10] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," in *SIGKDD Conf.*, 2007.

[11] F. Wei, W. Li, Q. Lu, and Y. He, "Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization," in *SIGIR Conf.*, 2008.

[12] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra Book and Solutions Manual.*   SIAM: Society for Industrial and Applied Mathematics, 2001.

[13] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22(1), 1951.

[14] E. M. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *VLDB Conf.*, 1998.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11(1), 2009.