An Effective and Economic Bi-level Approach to Ranking and Rating Spam Detection

Sihong Xie*, Qingbo Hu*, Jingyuan Zhang*, Philip S.Yu*
*Department of Computer Science University of Illinois at Chicago, Chicago, IL, USA

Abstract

Rating and ranking of items are important parts of modern electronic commerce. As a result, dishonest business owners are spamming the ecosystems in return for favorable product rankings, while consumers can be misled to purchase low quality products. To protect the interests of consumers, it is a critical task to spot spamming activities and maintain ecosystem health. Existing spam detection methods dichotomize the microscopic and macroscopic viewpoints of the problem. On the one hand, microscopic methods work on the scale of individual ratings and can be trapped in the ratings that are less harmful to the ecosystems health, leading to sub-optimal allocations of human efforts. On the other hand, macroscopic approaches focus on the ratings that can manipulate the ecosystems in a larger scale. However, the macroscopic signals they inspect can only be tangentially connected to the most critical system health statuses, leading to hard-to-measure spam detection outcome. Further, these macroscopic methods lack of a consistent way to drill down to the microscopic scale and detect actual spams. To address the above drawbacks, we propose a bi-level framework that unifies both perspectives to pinpoint suspicious ratings that can affect the ecosystems more directly and significantly, such that the limited human effort is allocated to maintain the ecosystem health effectively and economically. The framework revolves around the notion of ranking regularity. It first constructs a system health signal from an approximation of ground truth ranking via aggregation of multiple noisy crowdsourced rankings with minimal expert input. This signal helps the framework drill down on critical regions where a microscopic method can pinpoint suspicious individual ratings for human investigation. We obtain promising experimental results on datasets from mainstream restaurant rating websites.

1. Introduction

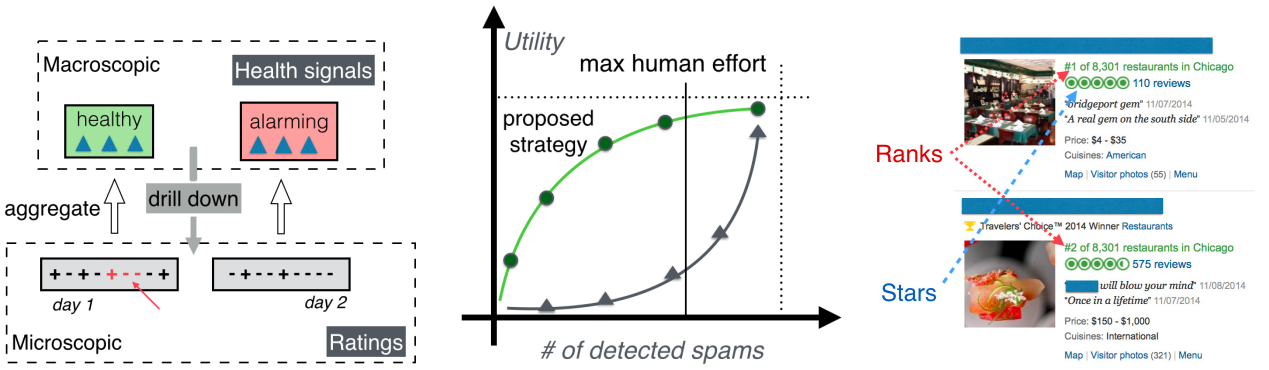
Nowadays, the collection of user-generated reviews and ratings on products and services has been an indispensable part of modern business, ranging from travel and dining to mobile apps. For example, customers of Amazon.com usually read reviews written by other customers before making purchase decisions. Clearly, business owners seek to have as many favorable reviews and ratings for their products as possible, and the fierce competition and huge revenue have driven some of the business owners to resort to underhanded rating/ranking manipulations. These manipulations can lead consumers to low quality products and services, causing unsatisfactory experience and even financial loss. It is therefore a critical task to detect such manipulated reviews/ratings and take proper measures to protect the consumers.

Spotting manipulations on ratings and reviews has been studied extensively, and existing approaches can roughly be divided into two categories, depending whether they take a microscopic or macroscopic view. Figure 1(a) shows these two different perspectives. Suppose ratings are grouped on a daily basis, and two such groups of ratings are shown at the bottom of the figure. From each group of ratings we derive signals indicating system health status (e.g. a fair and objective ranking of items), which are shown in the two boxes in the center. The first signal says the system is running healthily, while the second sounds an alarm. Microscopic methods work on the bottom (review or rating) level, and ignore the global health signals. In [16, 13, 12], they focus on characterizing fake individual reviews/reviewers/ratings via various behavioral and rating features. Though these methods are proved to be effective, the identified spams might only be tangentially correlated with the overall system health. As shown in Figure 1(a), without a global picture, the microscopic approach can find ratings (pointed by the red arrow) that do *not* actually cause the alarm. In practice, almost every *detected* review/rating has to be screened by human experts, which are limited and expensive resources, and it is unrealistic for every spam, detected or not, to get spotted and screened. As a result, these microscopic strategies can be inefficient as they spend the limited resources on the less critical spams and only improve the ecosystem health (utility) marginally, as shown by the green line in Figure 1(b).

Different from microscopic methods, macroscopic approaches aim at detecting high level signals (not involving individual rating or review) of ecosystem health. For example, the two Chicago restaurants topped on Tripadvisor list (Figure 1(c)) only received 4 stars and were not top-ranked on the other three popular rating websites. Such high level false/biased signals must be detected. In [33], they take the ranking perspective, and if an item is ranked too high in a short period among other items, it can be a suspicious one. In [31], they propose an algorithm to detect suspicious time windows where a business receives much more positive reviews than usual. However, these methods have the following drawbacks. First, these methods fail to define a sufficient indicator of health ecosystem status. For example, [31] adopts the signals of bursts in ranking or ratings of individual items, which are not directly connected to the overall ecosystem health (such as the product rankings that decide the order in which products are listed). Second, these methods are vague about how to spot the actual manipulating ratings/reviews that caused the detected abnormal system health statuses.

We propose a top-down bi-level framework that combines the strengths of both perspectives. The goal is to optimally

978-1-4673-8273-1/15/$31.00 ©2015 IEEE



(a) Flow chart of the proposed algorithm

(b) Human effort and utility curves of strategies

(c) Example of ranking and rating impression of restaurants

Figure 1: Motivations of the proposed framework

allocate a fixed budget of human effort to the most critical suspicious reviews/ratings. The number of spams is virtually beyond the capacity of human experts, and the screening procedure has to be stopped at a certain point, even there are remaining spams, and we want to stop at a point where the system utility is highest. The vertical line in Figure 1(b) indicates the cutting point when human effort reach its capacity, and the two curves represent two spam inspection strategies that end up with different system utilities. The proposed framework identifies suspicious reviews/ratings that lead to the fastest increasing in utility (ecosystem health). Overall, the utility of eliminating the identified spams should follow the "Law of Diminishing Marginal Utility", namely, after removing the most critical spams, the utility gained from removing the same amount of spams is decreasing until diminishes.

On the **macroscopic** level, we use ranking regularity to monitor system health status: given a system ranking over a set of items and the ground truth ranking, we can compute the correlation between the two rankings. A low correlation indicates that there can be certain manipulating activities such that the current system ranking significantly deviates from the true product ranking. On the one hand, putting a business on the top of a list is the fastest and most economic way [18] for a dishonest business to generate a large amount of revenues. On the other hand, if the products are in a reasonable order, then even there are some undetected spams, visitors are not severely affected, and laborious expert inspections on subtle reviews/ratings can be avoided. However, a single true ranking is usually hard to obtain. We address this issue via supervised aggregation of multiple less expensive crowdsourced rankings.

On the **microscopic** level the framework focuses on searching suspicious individual reviews and ratings. To save human efforts, this search is confined to the time windows where the system ranking is flagged in the macroscopic phase. Specifically, we adopt the Spearman's formula of ranking correlation, which measures how similar the items are ranked in two rankings. A low Spearman correlation indicates that there are items ranked much differently by the system ranking and the estimated ranking. We retrieve the top few items (businesses) that are farthest away in their positions in the two rankings, and for each identified suspicious item, we spot ratings that cause the large displacement in the two rankings of that item. Human efforts can then be saved on the remaining ratings as they either belong to items in good order or do not cause ranking discrepancy of a suspicious item. The proposed microscopic detection algorithm have parameters for controlling the number of detected spams, which is determined by the amount of available human effort. The contributions of the paper is summarized as follows:

- We propose a framework to unify the macroscopic and microscopic viewpoints to rating spams, with the capability of spotting spams that can directly and significantly affect online commerce ecosystems.
- We propose an effective macroscopic measurement of ranking system regularity. Though the implementation of such measurement requires a golden ranking of all items, we propose a more scalable and economic approach via supervised aggregation of multiple non-expensive rankings. We also propose various microscopic measurements to identify suspicious individual ratings for human screening.
- Experimental results on 3 real world review and ranking datasets demonstrate the effectiveness and feasibility of the proposed framework.

2. Preliminary

Traditional methods solve spam detection problem from either microscopic or macroscopic viewpoint. Microscopic methods mine certain characteristics of individual instances of review, rating or reviewer, such as text features, rating behaviors, etc. The goal of microscopic methods is to detect and eliminate suspicious instances. Though these methods can help clean up the ecosystem with the help of human experts, the characteristics of individual instances may not directly link to macroscopic system health status, such as fair rankings of items or objective opinions. As a result, it is difficult to tell whether one is inspecting the critical instances that are actually

Table 1: Notations

Symbol	Meaning
$U = \{1, \ldots, n\}$	Set of items to be ranked
n	Number of items to be ranked
π_0	Ground truth ranking, a permutation of U
$\hat{\pi}_0$	Estimated ground truth ranking
$\tau(\cdot, \cdot)$	Kendall-τ correlation between two rankings
$\rho(\cdot, \cdot)$	Spearman-ρ correlation between two rankings
$d(\cdot, \cdot)$	Distance between two rankings
K	Number of base rankings
$\pi_1, \ldots, \pi_K$	Base rankings
$\sigma_1, \ldots, \sigma_m$	Rankings to be examined
P	Markov transition matrix

manipulating the ecosystem, and how effective the current spam elimination is working. Human experts can spend the limited resources on the not-so-important spams and yet only scratch the surface of the macroscopic problem.

On the other hand, there are macroscopic approaches [33, 31] that directly investigate the overall health status of the rating/review system. By inspecting the system in a higher level with aggregated statistics, it is easier to tell whether the system is running healthily. More importantly, the limited human effort is only charged for investigating the spamming activities that are significantly affecting the ecosystem. As there are always undetected spams (or at least one cannot affirm the void of spams), while human effort is always limited, macroscopic methods can save the expensive human efforts for the most critical tasks. For example, in [31], they used aggregated rating in a time window to monitor rating of invidiaul items. If there is an alarm, their algorithm can drill down to more precise positions that contains suspicious activities, and human efforts are spent on a much smaller set of suspicious activities that affects item ratings. The drawback of this macroscopic approach is that the measurement they adopted (burst of ratings) may not be directly related to the more important ecosystem health status, such as the overall ranking of items. Indeed, ranking more directly determines the frequency of customers seeing an item. To resolve the issue, in [33], the authors proposed to use the rankings of items as health status of online stores. If an item's ranking is boosted frequently in a short period, this item is flagged as suspicious and call for human investigation. However, both methods lack of a microscopic view and cannot tell which individual ratings or reviews are suspicious, as a result, human experts can still be fuddled by a large volume of reviews/ratings.

3. The proposed bi-level framework

3.1. Macroscopic System health signals

There are various macroscopic signals of system health indicating whether the system is significantly affected by spamming activities. For example, the rank or average rating of a product can be computed as aggregated statistics, and if these statistics go up in a short period, there can be suspicious activities underlying the product ranking or rating. Though being effective [33, 31], these two signals have their drawbacks, respectively, and we argue that the ranking of *all* items is a more suitable and critical signal. First, the overall ranking of all items is directly connected to customer impression of the items. Therefore, product rankings are more likely to be targeted by spammers, and irregularities in ranking can directly affect customer decisions. Second, comparing to the work [33], ranking of all items is a system-wide status and tells much more of the story than the top-ranked individual products. Third, comparing to product rankings, it is more difficult to accurately measure the degree of irregularity of a burst in aggregated rating ("how bursty is bursty") [31]. Instead, we present a statistical rigorous approach to measure the irregularity of product rankings.

3.2. Ranking regularity via supervised ranking aggregation and ranking correlation

The regularity of overall ranking can be measured by the correlation between current system ranking to an authoritative ranking. However, there are two challenges. First, ranking all products in an objective and consistent way requires a lot of human efforts. For example, to score restaurants in a neighborhood, experts are paid to visit individual restaurants and then make their judgements. To ensure objectivity and freshness, a single restaurant has to be judged by multiple experts periodically, whose opinions are aggregated to obtain the final score. Such a procedure is expensive and cannot be scaled up to cover a wide range of restaurants. Second, one may resort to product ratings obtained from the crowd, where a large number of restaurants are rated from customers in an almost free way. However, such rankings are not good candidates for ground truth ranking, as previous studies [21, 16, 33] revealed that spams can be quite pervasive in crowdsourced ratings, and a single crowdsourced source can be too noisy to be helpful.

We propose to aggregate multiple crowdsourced rankings with a small amount of supervised information to address the above challenges. Existing studies show that aggregated rankings are more robust than base rankings [6]. Intuitively, the idea of ranking aggregation is similar to the averaging of multiple classifiers, and the averaged ranking can reduce bias and variance of base rankings, leading to a better ranking than the worst base ranking. However, obtaining an accurate ranking for effective anomaly detection requires more efforts. First, ordinary rank aggregations are unsupervised, and the resulting rankings can perform just slightly better than the base rankings, and the power of the constructed health detection signals can be quite limited. Second, it is possible to query a small amount of supervised information from the ground truth ranking. For example, a budget may be allocated for a group of experts to taste foods in a small number of restaurants, and a partial order can be made available. Such partial ground truth can be useful in guiding the ranking aggregation to produce a much better ranking. In fact, incorporating supervision was shown to be effective in boosting the power of rank aggrega-

tion [17]. In another perspective, supervised rank aggregation can be seen as using multiple base rankings to reduce the amount of supervision information needed for estimating the ground truth ranking. This is what we need, as it is less demanding to provide a small amount of pairwise product rankings than to ask for a total order of all products.

3.2.1. Ranking aggregation. Formally, given base rankings $\pi_1, \ldots, \pi_K$, a ranking aggregation algorithm produces a ranking $\hat{\pi}_0$ such that $\hat{\pi}_0$ satisfies certain optimality condition. We review some of the representative methods which also serve as the baselines in the experiments.

Heuristic approaches BordaCount and median rank aggregation [6] are based on positional ranks derived from multiple base rankings. Specifically, item i can be assigned a K dimensional vector $B(i) = (B_1(i), \ldots, B_K(i))$ where $B_k(i)$ is the number of entities ranked below i in the k-th base ranking. Then BordaCount assigns to the i-th entity the L_1 norm of $B(i)$ while median rank aggregation assigns to the entity the median of $B(i)$. Finally the aggregated ranking is obtained by sorting the items in descending order of the assigned positional ranks.

MLE based approaches One can assume that there is a latent variable that models the scores/ranks of the items, while the observed multiple rankings are generated by some generative model. Then the latent variables can be inferred using an MLE approach [5, 3, 2]. The Bradley-Terry (BT) model for rank aggregation is proposed in [3]. Let w_{ij} denotes the number of wins of item i to item j:

$$w_{ij} = \sum_{k=1}^{K} \mathbb{1}|\pi_k(i) < \pi_k(j)\}|$$

The log-likelihood of the BT model is

$$\ell(\lambda) = \sum_{1 \le i < j \le n} w_{ij}[\log \lambda_i - \log(\lambda_i + \lambda_j)] \tag{1}$$

where $\lambda = [\lambda_1, \ldots, \lambda_n]$ are the scores of the items and can be estimated using MLE. The aggregated ranking $\hat{\pi}_0$ is defined by ordering the items in descending order of $\lambda_i, 1 \le i \le n$. Note that the base rankings can be biased and affected by spammers, it is natural to down-weight the less reliable base rankings when inferring the underlying scores. For example, [5] extends the BT model to take care of the uncertainty in the base rankings.

Markov Chain based approaches This family of aggregation algorithms is closely related to the PageRank algorithm [6]. One can consider an item as a web page and the event that "item j is ranked higher than i" as "node i links to node j". Then a hyperlink network $G = (V, E)$ can be constructed. $V = \{1, \ldots, n\}$ is the set of nodes. If j is ranked higher than i, then $e_{ij} = 1$, and otherwise $e_{ij} = 0$. We can construct a Markov transition matrix for the k-th base ranking:

$$P_{ij}^{(k)} = \frac{e_{ij}}{\sum_l w_{il}} \tag{2}$$

Then the $n \times n$ matrix $P^{(k)}$ is just the transition matrix for a Markov random walk. The base rankings are aggregated by averaging their corresponding transition matrices:

$$P = \frac{1}{K} \sum_{k=1}^{K} P^{(k)} \tag{3}$$

The stationary distribution λ for $P^\top$ (i.e. $\lambda = P^\top \lambda$) is the "page rank" of the items.

3.2.2. Supervised Rank Aggregation. In order to incorporate any available supervision information to obtain a better estimation of ranking of items, we adopt supervised ranking aggregation. Formally, given base rankings $\pi_1, \ldots, \pi_K$, and $|L|$ pairs of ordered items, $L = \{\pi_0(i_1) < \pi_0(j_1), \ldots, \pi_0(i_{|L|}) < \pi_0(j_{|L|})\}$, a supervised ranking aggregation method produces a ranking $\hat{\pi}_0$ such that $\hat{\pi}_0$ satisfies certain optimality condition. Although there are some existing supervised ranking aggregation algorithms [17, 25, 1], we focus on an easy-to-implement yet effective algorithm. The proposed algorithm is based on the Markov Chain rank aggregation approach. We inject the supervision information into the transition matrix P to obtain $\tilde{P}$, as follows:

$$\tilde{P}_{ij} = \begin{cases} P_{ij} & \text{if } (i,j) \notin L, (j,i) \notin L, \\ 1 & \text{if } \pi_0(i) < \pi_0(j) \\ 0 & \text{if } \pi_0(i) > \pi_0(j) \end{cases} \tag{4}$$

The meaning of $\tilde{P}_{ij}$ is that if $(i,j) \in L$, then we are certain about the order of i and j and P_{ij} should be fixed as the supervised information, and otherwise should be estimated from the aggregated crowdsourced ranking. Next $\tilde{P}$ is normalized to obtain a column stochastic transition matrix, denoted by P with an abuse of notation. Finally, a ranking of the items can be obtained by computing the stationary distribution of $P^\top$.

3.2.3. Signal construction Algorithm. The procedure to construct the ranking-based health signal is described in Algorithm 1. Assume that we are monitoring the health of an online commerce ecosystem, where the ranking of items in the system can be summarized on a daily or weekly basis. The resulting daily or weekly rankings are denoted by $\sigma_1, \ldots, \sigma_m$. In step 5-7, we compute the stationary distribution (the PageRank of the items) using the power method. Each iteration of the power method takes $O(n^2)$ time for the product $P^\top \lambda$, which can be parallelized if n is huge. In step 10, we compute the correlation between the aggregated ranking $\hat{\pi}_0$ and each ranking of $\sigma_1, \ldots, \sigma_m$. These correlations serve as similarity measures, and if σ_i has a low correlation with $\hat{\pi}_0$, then σ_i is more likely to be an abnormal ranking of the items. There are two such measurements, namely, Kendall-τ

$$\tau(\pi_1, \pi_2) = \frac{2(c-d)}{n(n-1)} \tag{5}$$

and Spearman-ρ

$$\rho(\pi_1, \pi_2) = 1 - \frac{6 \sum_{i=1}^{n} (\pi_1(i) - \pi_2(i))^2}{n(n^2-1)} \tag{6}$$

where c is the number of pairs of items that are concordant/discordant (in the same/reverse order) in two rankings π_1 and π_2. The labels of the test product rankings can be obtained by thresholding the correlations between the test rankings and the ground truth ranking (step 8). The transformation that follows computes the probability of σ_i being an abnormal ranking $p(y_i = 1|\sigma_i)$ (step 9-11). Note that although we describe the input of the algorithm to be a batch of system rankings, it is clear that the current health signal $p(y_i = 1|\sigma_i)$ can be obtained on the fly as the test rankings σ_i stream in. This is a desirable property for the algorithm to deployed as a monitoring scheme.

Algorithm 1 Ranking regularity based ecosystem health monitoring

1: **Input**: Base rankings $\pi_1, \ldots, \pi_K$, gold pairwise rankings $L = \{\pi_0(i_1) < \pi_0(j_1)), \ldots, \pi_0(i_{|L|}) < \pi_0(j_{|L|})\}$, system rankings $\{\sigma_1, \ldots, \sigma_m\}$
2: **Output**: Estimated ranking irregularities $\hat{p}(y_i = 1|\sigma_i), i = 1, \ldots, m$
3: Construct the transition matrix P of the Markov chain as in Eq.(4), using $\pi_1, \ldots, \pi_K$ and L.
4: $\lambda = \text{rand}(n, 1)$. # *random vector of length* n
5: **while** not convergent **do**
6: $\quad \lambda = P^\top \lambda$
7: **end while**
8: Estimate the ground truth ranking π_0 by ordering λ.
9: **for** Each σ_i **do**
10: $\quad s_i =$Kendall-$\tau(\sigma_i, \hat{\pi}_0)$ or Spearman-$\rho(\sigma_i, \hat{\pi}_0)$.
11: $\quad \hat{p}(y_i = 1|\sigma_i) = 1/(1 + \exp(s_i))$
12: **end for**

3.3. Microscopic detection of manipulating ratings/reviews

Having located the major regions (such as certain time windows) where the system rankings deviate significantly from the normal ranking, we propose a procedure to drill down to a microscopic level and identify suspicious individual ratings. Suppose σ is an instance of system ranking that has small correlation with the aggregated ranking $\hat{\pi}_0$. Our goal is to investigate any suspicious ratings that possibly cause the detected ranking irregularities. We accomplish this goal in two steps: firstly we find suspicious businesses that have irregular ranks, and secondly we identify individual ratings that are more likely contribute to such ranking irregularities.

In the first step, we drill down from the *macroscopic* ranking irregularity to the abnormal rankings of individual businesses via Eq.(6). Specifically, if the ranking correlation $\rho(\sigma, \hat{\pi}_0)$ is too low, then the second term in Eq.(6) has a large value. This large value can be caused by any of the summands $(\sigma(i) - \hat{\pi}_0(i))^2$ in the numerator, which is the square of the distance in ranking positions of a business in the two rankings. Therefore, one should focus on the business that have such large distances. The businesses that have close ranking positions are less harmful as they contribute less to the low ranking correlation. In the second step, we further drill down to individual ratings for the identified suspicious businesses, by identifying those ratings that contribute to the large displacements of the businesses. In particular, the ranking irregularity of a business can either be caused by overrating ($\sigma(i) \ll \hat{\pi}_0(i)$) or underrating ($\sigma(i) \gg \hat{\pi}_0(i)$). Therefore, we flag ratings that either overrate or underrate that business accordingly. The algorithm is described in Algorithm 2. *Remarks:* the parameters k and ℓ control the number of suspicious ratings to be retrieved, and the larger these two values are, the more suspicious ratings will be reported, making the algorithm more sensitive to spams. In practice, one can start from smaller values for retrieving the most suspicious ratings for investigation. If time permits or the system ranking still look abnormal, larger values can be used to find more suspicious ratings.

The microscopic detection step described above distinguishes the proposed framework from previous works [33, 31]. In [33], they only identify suspicious businesses or items using ranking, without identifying suspicious individual rating. In [31], they identify suspicious extremely high ratings, which are however not explicitly connected to anomalies in the rankings of businesses. The proposed framework addresses these shortcomings by starting from macroscopic ranking anomalies and drilling down to suspicious businesses and ratings on the microscopic level.

Algorithm 2 Microscopic Suspicious Rating Identification

1: **Input**: system ranking of items σ, estimation of ground truth ranking $\hat{\pi}_0$. Parameters k and ℓ.
2: **Output**: identified suspicious ratings S.
3: **for** Each business i **do**
4: $\quad$ Compute the displacement between the ranking positions of i: $d_i = (\sigma(i) - \hat{\pi}(i))^2$.
5: **end for**
6: $S = \{\}$.
7: **for** Each business i in top k ones that have the largest d_i **do**
8: $\quad$ **if** $\sigma(i) \gg \hat{\pi}_0(i)$ **then**
9: $\quad\quad S = S \cup$ the highest ℓ ratings for business i.
10: $\quad$ **else if** $\sigma(i) \ll \hat{\pi}_0(i)$ **then**
11: $\quad\quad S = S \cup$ the lowest ℓ ratings for business i.
12: $\quad$ **end if**
13: **end for**

4. Experiments

4.1. Experimental settings

Datasets We employ rating data of restaurants in several areas, including New York City (NYC), Chicago (CHI), Phoenix (PHX) in the experiments. The input of the framework requires multiple product rankings, which are collected from 3 popular rating websites: tripadvisor.com, yelp.com and foursquare.com. These websites summarize user ratings to assign a single score to each restaurant. We transform these scores to restaurant rankings and thus the difference in the scales of rating systems does not matter.

Table 2: Characteristics of Data

	NYC	CHI	PHX
# of restaurants	79	76	77
# of ratings	12733	11175	8381
# of σ_i (m)	275	254	208
# of positive inst	90	37	94

We also need a ground truth ranking to provide supervision information. Similar to the experiments in [23], we adopt the restaurant ranking from Zagat.com as ground truth. Different from the crowdsourced rankings from Tripadvisor, Yelp and Foursquare, Zagat has a more reliable and consistent way of restaurants evaluations. The validity of the Zagat ranking is also recognized by New York Times as "a necessity second only to a valid credit card" [1].

4.2. Macroscopic irregularities detection

On the macroscopic level, our goal is to identify time windows where the system ranking deviates significantly from the ground truth ranking. We crawl the ratings of the restaurants in the three areas from tripadvisor.com to evaluate the macroscopic detection performance. The ratings are aggregated on a weekly basis, and we order the restaurants based on the average rating each restaurant receives in each week. The resulting weekly restaurant rankings, denoted by $\{\sigma_1, \ldots, \sigma_m\}$, are assigned ground truth labels as follows: we compute the Kendall-$\tau(\sigma_i, \pi_0)$ or Spearman-$\rho(\sigma_i, \pi_0)$ where π_0 is the ranking from Zagat.com, and assign 1 (abnormal) to σ_i if the correlation is less than 0 (negatively correlated), and -1 (normal) otherwise. In this way, we created two sets of labeled rankings with two ranking correlation measures. The characteristics of the datasets are summarized in Table 2. Note that although we have only hundreds of rankings for detection, there can be tens of thousands of ratings that can potentially be spams and need to be inspected by human beings.

Baselines We have two groups of baselines for ranking anomaly detection. The first group of baselines use a single crowdsourced ranking for the detection. Specifically, we have rankings of restaurants from yelp (YLP), foursquare (FQ), tripadvisor (TA). The second group of baselines use aggregated rankings (BT, CrowdBT, and BordaCount (BC)) for the detection, as described in Section 3.2.1. The implementation of BordaCount is straightforward. For BT, we adopt the R package "BradleyTerry2" [2]. We implement CrowdBT according to [5]. To make predictions using these single or aggregated rankings, we follow step 10 of Algorithm 1 and construct macroscopic detection signals for $\sigma_1, \ldots, \sigma_m$, but with the estimated ranking $\hat{\pi}_0$ set to the rankings obtained from the baselines. Note that we have two different ranking correlation measurements, thus when computing these signals, Kendall-τ or Spearman-ρ is employed according to which measure is used to assign ground truth labels. The performance of the proposed algorithm depends on the amount of supervision provided, which varies from 20 to 200 pairs of randomly selected gold pairwise rankings. We report the averaged performance over 100 repeats of the experiments.

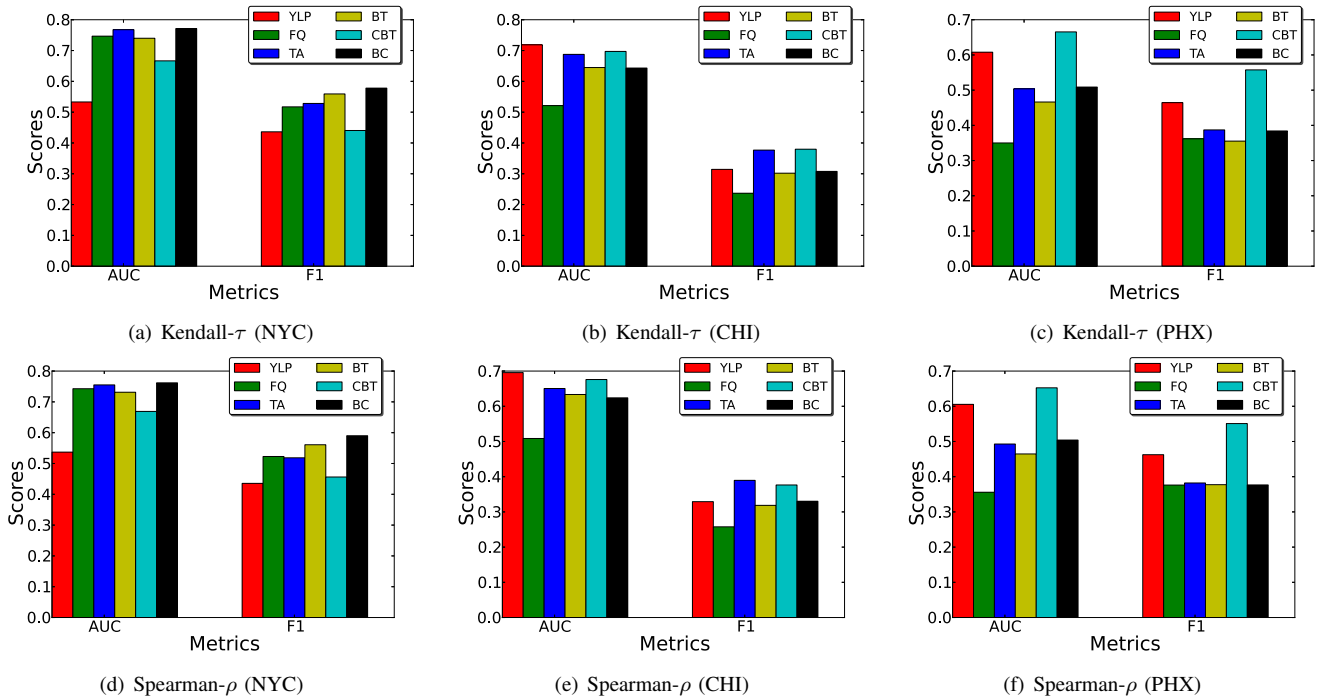
Macroscopic Performance Metrics We adopt AUC and F1 score as our metrics, as these are standard for evaluation in anomaly detection situations. AUC is the area under the FP-TP curve, and the higher the better. F1 score depends on two other metrics: recall and precision rates. Recall rate is the percentage of discovered positive instances among all positives (some of them may not be discovered), while precision rate is the percentage of discovered positives among all discovered instances (could include negatives). Intuitively, an anomaly should be "discovered" for human inspection and we would like the detection algorithm to assign a high score to a true anomaly. Higher recall and precision would be more desirable. However, one cannot expect both metrics to be high at the same time as there is a trade-off between them. F1 score considers this situation and takes the harmonic mean of recall and precision as a single metric. To compute F1 score (also recall and precision), one needs to predict the labels of weekly rankings. We assign label +1 (-1, resp.) to a weekly ranking if its correlation with the estimated ranking $\hat{\pi}_0$ is less (greater, resp.) than 0.

4.2.1. Macroscopic Performance. Comparing the baselines The performance of two groups of baselines are plotted in Figure 2 (best viewed in color). The figures on the first (second, resp.) row use Kendall-τ (Spearman-ρ, resp.) as ranking correlation measurement. We order the bars in each subfigure such that the first three correspond to the first group of baselines (single crowdsourced rankings) and the latter three correspond to the second group (*unsupervised* ranking aggregations). From the figures, we have the following observations. First, in all cases except the AUC metric in Figure 2(b), the best baseline from the second group outperforms the best one from the first group. For example, from the two figures in the first column on the NYC dataset, BC (the black bars) is the best method in all six baselines, under both ranking correlation measurements. This observation confirms that aggregating multiple biased and noisy base rankings can be better than any single base ranking. Second, there is no obvious winner in the second group. For example, in the two figures in the last column on the PHX dataset, CBT, instead of BC, outperforms the other methods with a large margin in both ranking correlation measurements. In practice, there are many cities and how to pick the right aggregation method becomes an issue. This observation justifies the proposed detection signal based on *supervised* rank aggregation.

Comparing the bi-level framework with the best baselines In Figures 3(a) $\sim$ 3(l), we compare the macroscopic detection performance of the proposed framework (blue lines) to the best ones from the two groups of baselines (green and red lines), respectively. The first two rows of figures are obtained using Kendall-τ as ranking correlation measure,

1. http://www.nytimes.com/1997/11/12/dining/zagat-s-new-york-survey-entries-up-contributors-off.html

2. http://cran.r-project.org/package=BradleyTerry2



(a) Kendall-τ (NYC) (b) Kendall-τ (CHI) (c) Kendall-τ (PHX)

(d) Spearman-ρ (NYC) (e) Spearman-ρ (CHI) (f) Spearman-ρ (PHX)

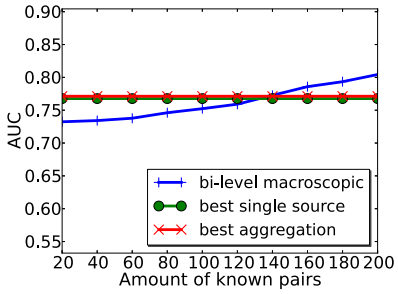
Figure 2: Baseline Comparisons

and the latter two rows are obtained using Spearman-ρ. For each ranking correlation metric, we plot the AUC and F-1 performance metrics. We have the following observations. First, with a small amount of supervision, the proposed method outperforms all the six baselines. For example, on the CHI dataset, only 40 pairs of gold pairwise rankings is sufficient for an equally good AUC of the best baseline. To put this number into perspective, one only needs to taste foods in $\sqrt{40} \approx 7$ restaurants to obtain this supervision information, as one can transform $O(n)$ scores into $O(n^2)$ pairwise rankings. Second, with more supervision, the performance of the proposed bi-level method goes up steadily in all metrics. This confirms the desirable consistency of the proposed algorithm: so long as the budget allows, superior detection performance can be obtained almost surely.

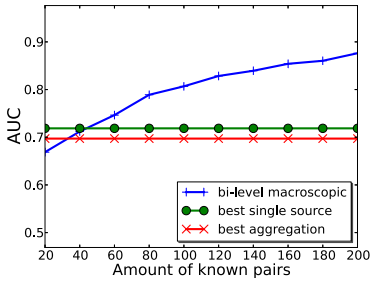
4.3. Microscopic performance study

4.3.1. A case study. Previous studies [31, 27, 16] show that it is a difficult task to conduct human evaluation on a large set of suspicious reviews. Also in [13], they found it difficult for human being to judge whether a review is a dishonest one or not. Therefore, we demonstrate that on the microscopic level, the proposed framework can find interesting traits of spammers, providing cues as starting point for human evaluation and input to other spam detection algorithms. The upper part of Figure 4 shows a spotted review posted on Aug 21, 2008. The proposed framework spot this review since the system ranking in the week of Aug 21, 2008 has negative ranking correlation with the estimated ground truth ranking, and the business for which this review is posted has the largest distance in the two rankings. Specifically, among all 6 businesses, this business is displaced 4 places higher up, which indicates overrating. By looking at the first review, it might be hard to judge its nature. However, we check the reviews written by the same account, and find out that, as shown in the same figure, the same account posted many reviews on the day (Jun 30, 2008) which is a strong evidence of spamming. This example shows that the proposed framework can find reviews that lead to the detection of more reviews during human evaluation.

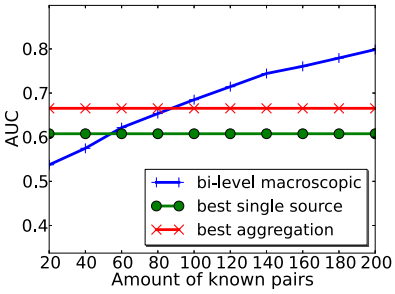
4.3.2. Sensitivity of parameters. We also demonstrate how the total number of ratings reported by the framework varies as the threshold of ranking correlation measurements takes different values over the 6 year period. Here we use the Spearman-ρ as the ranking correlation. The number of reported ratings is divided into two parts, namely, ratings that lead to overrated and underrated businesses. The results are shown in Figure 5. As can be seen from the figure, as the threshold goes up, more and more weekly rankings will be flagged as "far away" from the estimated ground truth ranking in the macroscopic phase, leading to more ratings flagged as suspicious in the microscopic phase. Note that although for each city, there are about 1,000 reviews that need human screening for all three cities over the 6 year period (about 3 reviews to screen per week), yet consider that there are tens of such major cities in the world, the workload of screening can be understated. However, this number is already a huge reduction ($< 0.33\%$) from the total number (32,289) of all reviews in our dataset.



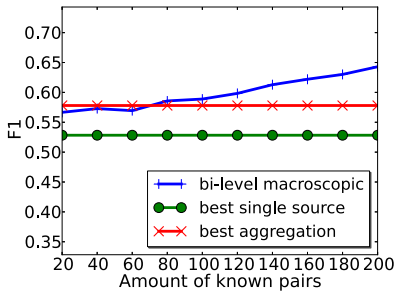
(a) AUC in Kendall-τ (NYC)



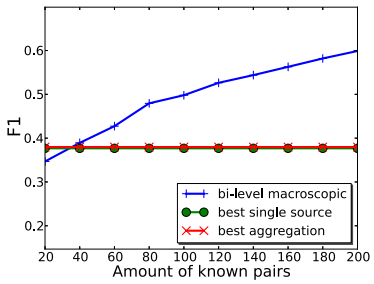
(b) AUC in Kendall-τ (CHI)



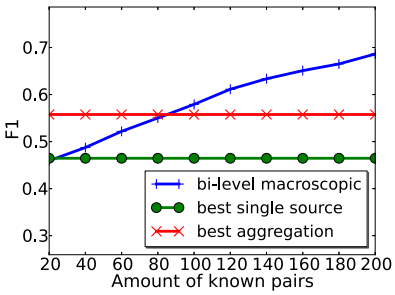
(c) AUC in Kendall-τ (PHX)



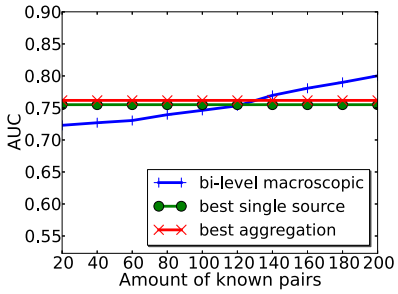
(d) F1-score in Kendall-τ (NYC)



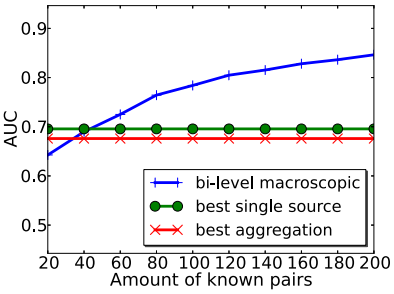
(e) F1-score in Kendall-τ (CHI)



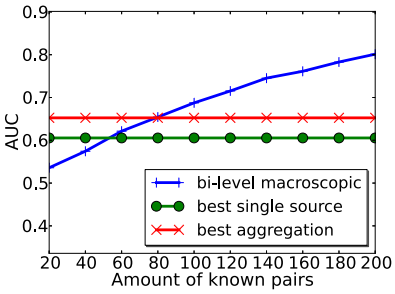
(f) F1-score in Kendall-τ (PHX)



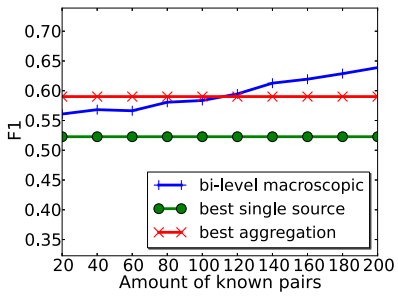
(g) AUC in Spearman-ρ (NYC)



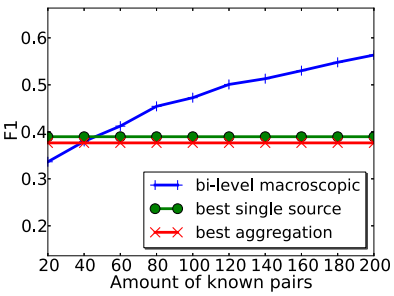
(h) AUC in Spearman-ρ (CHI)



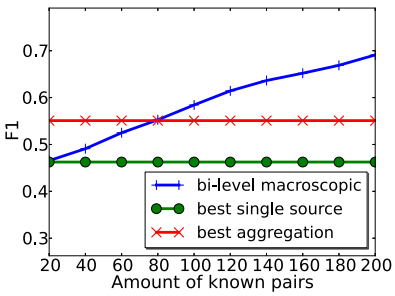
(i) AUC in Spearman-ρ (PHX)



(j) F1-score in Spearman-ρ (NYC)



(k) F1-score in Spearman-ρ (CHI)



(l) F1-score in Spearman-ρ (PHX)

Figure 3: Macroscopic performance comparison

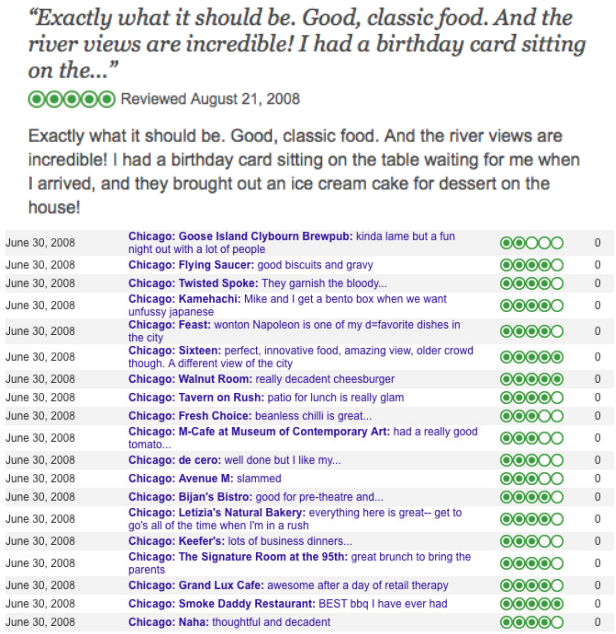


Figure 4: Example of spotted spams

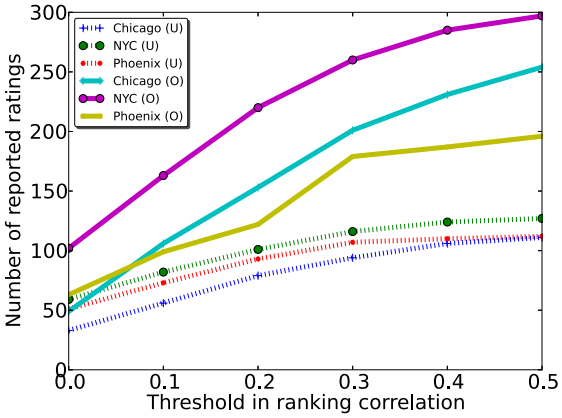


Figure 5: Number of reported suspicious ratings

5. Related Work

Ranking spam detection Ranking spams refer to any suspicious activities on the web to promote or demote any business, product, services, etc. Typical examples include writing undeserved positive or negative reviews, giving very high ratings in a short period to restaurants, or create webpages with many links to the promoted webpage to cheat search engines, etc. [31, 16, 13] study spams that give undeserved high ratings to products, where the suspicious objects are reviews, reviewers, or products, but they do not study the credibility of the whole ranking of products. [27] uses a tri-partite graph to encode the relationship between businesses, reviewers and reviews, and then suspicious scores of the three types of objects can be derived using a random walk like algorithm. In this view, their work also focus on the credibility of individual objects instead of the whole ranking. [21] detects groups of reviewers who write similar reviews, which have nothing to do with rankings of businesses. In [19], they study blog spams and propose an algorithm to block them using language model. [14, 27, 30] study the detection of spams in comments that reply to blogs or news articles. We have a different point of view of suspicious activities from the above existing works.

Rating calibration and estimation try to reveal the true ratings of the items. In [15], they propose a reinforcement framework to jointly infer the bias of individual raters and the controversy of ratings in each items. The quality of the ratings can also be inferred as a by-product. In [20], they try to simultaneously find the unbiased comment ratings and user bias using a reinforcement framework similar to that in [15]. Interestingly, they also find that by incorporating labeled information in their framework, the performance can be substantially boosted. In [4], they propose a hierarchical model to model rater reputations and the quality of comments. Their method depends on textual features of the comments, and thus is not applicable to our case. Overall, the above methods can identify "good" and "bad" items, but do not provide a macroscopic view of ranking systems. In [5], they study the problem of finding "reputable" content sharers, who are more likely to share high quality content. In [24], they propose a framework to model and predict the "helpfulness" of reviews. Similarly, the above work focus on the quality of individual items and do not take the macroscopic perspective. In [32], they focus on the problem of co-ranking contents and users in Q&A sites. Their work does not study the bad reviews, and thus is not comparable to our work here. The authors in [26] focus on the information propagation perspective of ranking and try to leverage the similarities of researcher works to adjust the ranking of the authority of the scholars. In [9, 10], the authors studied the quality of online videos in terms of "longevity", which is estimated based on time series of viewer activities and can be considered as a "crowdsourced ranking".

Rank aggregation is an important research subject in machine learning and information retrieval. For example, in meta-search, where multiple search engines can cover different portions of the web while an end-user usually has multiple criteria his/her search [1]. Two models that extends the BT model is the Plackett-Luce model and Thurstonian model. [11] provides an EM-like algorithm to infer the generalized BT model. [8] models the rank distribution from a Bayesian perspective and uses an expectation propagation to infer the model. [29] considers the online inference of ranking. There are some learning theoretical research on rank aggregation. In [28], they study the sample complexity of rank aggregation, namely, how many observations are needed to recover the true ranking with high probability. In [22] studies the convergence of various rank aggregation, including the BT model, rank centrality model and BordaCount model. In [7] propose to use matrix completion to handle the noise and incomplete information in rank aggregation.

6. Conclusion

In this paper we propose a bi-level framework to effectively and economically identify suspicious ratings that are more likely manipulating the ecosystem's health, such as overall ranking and rating of the products. To save the limited human experts' resources to the most critical suspicious ratings and reviews, the framework combines the macroscopic and microscopic viewpoints that were dichotomized by existing spam detection methods, and spots suspicious ratings by drilling down from higher level of ecosystem health status to the lower level of individual business and ratings. Experimental results shows that the proposed framework is promising.

Acknowledgment

This work is supported in part by NSF through grants III-1526499, CNS-1115234, and OISE-1129076, Google Research Award, and the Pinnacle Lab at Singapore Management University.

References

[1] Javed A. Aslam and Mark Montague. Models for metasearch. SIGIR, 2001.

[2] Ralph Allan Bradley. Rank analysis of incomplete block designs: Ii. additional tables for the method of paired comparisons. *Biometrika*, 41(3/4):pp. 502–537, 1954.

[3] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):pp. 324–345, 1952.

[4] Bee-Chung Chen, Jian Guo, Belle Tseng, and Jie Yang. User reputation in a comment rating environment. KDD, 2011.

[5] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. WSDM, 2013.

[6] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. WWW, 2001.

[7] David F. Gleich and Lek-heng Lim. Rank aggregation via nuclear norm minimization. KDD, 2011.

[8] John Guiver and Edward Snelson. Bayesian inference for plackett-luce ranking models. ICML, 2009.

[9] Qingbo Hu, Guan Wang, and Philip S. Yu. Deriving latent social impulses to determine longevous videos. WWW, 2014.

[10] Qingbo Hu, Guan Wang, and P.S. Yu. Assessing the longevity of online videos: A new insight of a video's quality. In *DSAA*, 2014.

[11] David R. Hunter. Mm algorithms for generalized bradley-terry models. *The Annals of Statistics*, 2004.

[12] N. Jindal and Bing Liu. Analyzing and detecting review spam. ICDM, 2007.

[13] Nitin Jindal and Bing Liu. Opinion spam and analysis. WSDM, 2008.

[14] Ravi Kant, Srinivasan H Sengamedu, and Krishnan Kumar. Comment spam detection by sequence mining. WSDM, 2012.

[15] Hady W. Lauw, Ee-Peng Lim, and Ke Wang. Bias and controversy: Beyond the statistical deviation. KDD, 2006.

[16] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. CIKM, 2010.

[17] Yu-Ting Liu, Tie-Yan Liu, Tao Qin, Zhi-Ming Ma, and Hang Li. Supervised rank aggregation. WWW, 2007.

[18] Luca M. Reviews,reputation,andrevenue:the case of yelp.com. In *Harvard business school working papers, Harvard Business School*, 2011.

[19] Gilad Mishne, David Carmel, and Ronny Lempel. Blocking blog spam with language model disagreement. In *AIRWeb*, 2005.

[20] Abhinav Mishra and Rajeev Rastogi. Semi-supervised correction of biased comment ratings. WWW, 2012.

[21] Arjun Mukherjee, Bing Liu, and Natalie Glance. Spotting fake reviewer groups in consumer reviews. WWW, 2012.

[22] Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *ICML*, 2014.

[23] Chenhao Tan, Ed H. Chi, David Huffaker, Gueorgi Kossinets, and Alexander J. Smola. Instant foodie: Predicting expert ratings from grassroots. CIKM, 2013.

[24] Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. Context-aware review helpfulness rating prediction. RecSys, 2013.

[25] Maksims N. Volkovs and Richard S. Zemel. Crf framework for supervised preference aggregation. CIKM, 2013.

[26] Guan Wang, Qingbo Hu, and Philip S. Yu. Influence and similarity on heterogeneous networks. CIKM, 2012.

[27] Guan Wang, Sihong Xie, Bing Liu, and Philip S. Yu. Identify online store review spammers via social review graph. *ACM Trans. Intell. Syst. Technol.*, 2012.

[28] Fabian L Wauthier, Michael I Jordan, and Nebojsa Jojic. Efficient ranking from pairwise comparisons. ICML, 2013.

[29] Ruby C. Weng and Chih-Jen Lin. A bayesian approximation method for online ranking. *J. Mach. Learn. Res.*, 12, 2011.

[30] Benny Wong, Michael E. Locasto, and Angelos D. Keromytis. Palprotect: A collaborative security approach to comment spam. IEEE Information Assurance Workshop, 2006.

[31] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S. Yu. Review spam detection via temporal pattern discovery. KDD, 2012.

[32] Jingyuan Zhang, Xiangnan Kong, Roger Jie Luo, Yi Chang, and Philip S Yu. Ncr: A scalable network-based approach to co-ranking in question-and-answer sites. In *CIKM*, 2014.

[33] Hengshu Zhu, Hui Xiong, Yong Ge, and Enhong Chen. Ranking fraud detection for mobile apps: A holistic view. CIKM, 2013.