Robust Crowd Bias Correction via Dual Knowledge Transfer from Multiple Overlapping Sources

Sihong Xie*, Qingbo Hu*, Jingyuan Zhang*, Jing Gao†, Wei Fan‡, Philip S.Yu*
*Department of Computer Science University of Illinois at Chicago, Chicago, IL, USA
†Department of Computer Science, University at Buffalo, Buffalo, NY, USA
‡Baidu Research Big Data Lab, Sunnyvale, CA, USA

Abstract—**One of the largest constituents of big data is the crowdsourced or user-generated data which contain a wide range of valuable information. However, they are inherently biased and possibly spammed, making trustworthy information extraction an imperative task. As a special case, we study reviewer-posted ratings for products. The biased ratings can lead to disappointed customers due to overrated products, and reduced revenues of business owners caused by undeserved negative ratings. To distill objective product quality measurements, most existing methods try to infer unbiased ratings from the raw ratings alone, and may not overcome the inherent bias to recover the underlying true ratings. Though improved bias corrections have been achieved with domain expert helps, the overhead of expert efforts can be rather expensive in practice. We exploit the variety of big data and adopt a multiple source mining approach, which finds trustworthy measurements without domain expert, but with knowledge crowdsourced and transferred from external domains. We address the challenges that the multiple data sources are 1) inherently heterogeneous, 2) at most only partially overlapping and 3) biased by themselves. We explore and analyze the strengths and weaknesses of various knowledge transfer strategies. We then propose Consensus Ranking Dual Transfer (CRDT) to handle the above challenges by identifying "anchor reviewers" as a bridge for robust "dual transfer", and removing bias in individual sources via consensus ranking aggregation. Experiments on real-world rating datasets demonstrate that the proposed approach can deliver more robust bias correcting effects than the baselines and can identify abnormal reviewers.**

I. Introduction

Crowdsourced and user-generated data are an important part of the big data being accumulated, such as the prevalent product rating and review data: hotels worldwide are reviewed and rated on Tripadvisor; several hundred million products are rated and reviewed by customers on Amazon. By hosting such comprehensive and opinionated data, these systems are not only vital to customers, but also to business owners. However, bias and noises are inevitable in these crowdsourced data, and to make sense of the data, it is important to infer objective and fair product quality measurements, such that customers can make informed decisions and business owners are not hurt by undeserved negative ratings.

The task is non-trivial since the crowdsourced data are biased for several reasons. First, although most of these rating systems employ state-of-the-art quality control mechanisms such as ReCaptcha [19], spammers can still infiltrate the systems and give arbitrary ratings to entice customers into purchasing of low quality products. Second, affected by uncontrollable factors, regular users may rate products subjectively or spontaneously. For example, the perceptions of the quality and the ratings of a product can vary dramatically among reviewers.

There are existing works trying to address these issues in the pursue of objective product quality measurements. In [10, 9], they proposed an unsupervised method to jointly infer reviewer bias and product quality. These methods did not exploit supervision information and can possibly be misled by the inherent rating bias. In [16], they proposed to incorporate ground truth ratings to achieve better results. They further adopt active learning to further reduce the cost of expert supervisions. In [18], the authors proposed a supervised matrix factorization model to infer multiple latent factors, based on which expert ratings can be predicted using crowdsourced ratings. However, their method also require a significant amount of supervisions. In general, supervised methods are more effective in recovering unbiased information, but can incur expensive expert efforts.

Fortunately, big data provide a rich set of data sources, which we exploit via transfer learning to avoid the expensive expert efforts. The basic assumption is that multiple data sources contain information of the same subset of products. When expert ratings are too expensive to collect, weak supervisions transferred from auxiliary sources can be used to substitute expert input. For example, IMDB and Netflix can provide useful information about movies to infer a less biased movie ranking on Amazon.

We identify the following challenges in trading domain experts for multiple auxiliary sources. First, being user-generated, it is common for product ratings/rankings from multiple auxiliary domains to be inconsistent or conflicting. For instance, IMDB might give 5 stars to a movie while Netflix gave the same product 3 stars. The challenge is to resolve such conflicts across domains. Second, due to the difference among domains, it is likely that the auxiliary domains only partially overlap with the target domain, and thus are not directly helpful to those non-overlapping products.

Table I
NOTATIONS

Symbol	Meaning
$\mathcal{U}$	Set of users/reviewers, or the crowd
$\mathcal{V}$	Set of products to be rated
$\mathcal{R}$	Ratings of products from the users/reviewers
$n = \|\mathcal{V}\|$	Number of products
$m = \|\mathcal{U}\|$	Number of users
q_j	Quality of the j-th product
b_i	Bias of the i-th user
$\pi_1, \ldots, \pi_K$	Partial orderings of $\mathcal{V}$
π_0	Ground truth ranking of $\mathcal{V}$
σ	Score function of $\mathcal{V}$
$\tau(\cdot, \cdot)$	Kendall-τ ranking correlation coefficient
$\rho(\cdot, \cdot)$	Spearman-ρ ranking correlation coefficient
L	Set of pairwise ranking constraints on products
S	Set of score constraints on products

Lastly, the auxiliary sources can be biased themselves, and are not immediately useful to the target domain.

We address the above challenges by a two-step pipeline. The first step is to resolve inter-source conflicts. It cancels out the bias in individual auxiliary sources, and extracts a single consensus product rating/ranking as transferable knowledge. The second step is to transfer knowledge from auxiliary source to the target domain. After exploring two product-centric transfer learning strategies, which are less effective in the presence of non-overlapping products, we propose a dual transfer approach, which applies the transferred knowledge to both products and reviewers (thus the name "dual"). Both the reliability of anchor reviewers and product ranking are estimated using auxiliary sources to regulate the bias correction procedure. Experiments on three real-world datasets show that the dual transfer approach outperforms previous approaches and the two single transfer approaches. Furthermore, we show that the inferred bias can be used as a signal for suspicious reviewer identification.

II. PRELIMINARY

Suppose we have n products $\mathcal{V} = \{v_1, \ldots, v_n\}$ which are rated by m reviewers $\mathcal{U} = \{u_1, \ldots, u_m\}$. Let r_{ij} denote the rating given by user u_i to product v_j, and $\mathcal{R}$ denote the collection of all ratings. $\mathcal{R}$ can be seen as the union of $\mathcal{R}^j$ (ratings dedicated to the j-th product): $\mathcal{R} = \cup_j \mathcal{R}^j$, or the union of $\mathcal{R}_i$ (ratings given by the i-th reviewer): $\mathcal{R} = \cup_i \mathcal{R}_i$.

A ranking of the products is a function $\pi : \mathcal{V} \rightarrow \{1, \ldots, n\}$, and $\pi(v_i) < \pi(v_j)$ (or $v_i \succeq_\pi v_j$) means that product v_i ranks higher (is better) than v_j. Let π_0 denote the unknown ground truth product ranking, the querying of which is expensive. Our goal is to correct bias in $\mathcal{R}$, such that the estimated ranking $\hat{\pi}$ is as close to π_0 as possible. The notations are summarized in Table I.

A. Unsupervised bias correction

A representative method is proposed in [16], which tries to infer unbiased product quality soly from crowdsourced ratings. Associate the product v_j with a quality score q_j and reviewer u_i with bias b_i. We impose a reinforcement relationship between the product quality and reviewer bias:

$$q_j = \frac{1}{|\mathcal{R}^j|} \sum_{i \to j} r_{ij}(1 - b_i) \qquad (1)$$

$$b_i = \frac{1}{|\mathcal{R}_i|} \sum_{i \to j} |r_{ij} - q_j| \qquad (2)$$

The quality of a product is the averaged ratings dedicated to that product, adjusted by individual user bias, and the bias of a user is the averaged distance from his/her ratings to the estimated quality of the products he/she has rated.

Unfortunately, as pointed out in [16], without expert guidance, the above unsupervised algorithm is not very effective in inferring the true product quality. Suppose that the majority of the ratings for a product are fake ratings and biased toward the highest score, say 5-star, then during the first iteration, the estimated rating of the product is the average of the biased ratings and will be seriously biased towards 5-star. When a dishonest reviewer has only give a 5-star rating to that product, then the bias of this reviewer will be low, since his/her only rating agrees with the dominating fake 5-star ratings of the product.

B. Semi-supervised bias correction

In [16], the authors proposed to clip the scores of a subset of the products to expert evaluations, and infer the remaining product scores using the following equation:

$$q_j = \begin{cases} S(j) & \text{if } j \in S \\ \frac{1}{|\mathcal{R}^j|} \sum_{i \to j} r_{ij}(1 - b_i) & \text{otherwise} \end{cases} \qquad (3)$$

where S is the set of products whose scores that are fixed at their ground truth scores (denoted by $S(j), j \in S$), and the reviewer bias is estimated as in Eq. (2). The above equations propagate the supervision information about S to the remaining reviewers and products. By incorporating expert ratings, the method can better correct the rating bias, and the restored product scores are closer to editorial ratings. However, this semi-supervised algorithm requires large amount of input from experts ([16] labeled 50% of the products) to counteract the sensitivity of the graph-based propagation algorithm to the labeled set S.

III. CORRECTING CROWD BIAS VIA TRANSFER LEARNING

We assume that there is no expert input to guide the bias correction, and seek for help from related external domains. We first explored two product-centric single transfer strategies, and then we point out their drawbacks and propose a dual transfer strategy.

A. Two product-centric single transfer strategies

1) Product rating single transfer: One can alter the semi-supervision bias correction algorithm in Section II-B, and choose the overlapping products that are rated in both target and auxiliary domains as the set S. The ratings in S are fixed to the averaged scores computed from multiple auxiliary domains, while the ratings of the non-overlapping products are estimated as in Eq. (1). Compared to the unsupervised bias correction, the transferred ratings are the average of several auxiliary domains. The underlying assumption is that the aggregated information from multiple data sources can be potentially less biased and be further propagated to the non-overlapping products, thereby overcoming the dominance of the biased ratings in the target domain. However, different domains typically have different rating scales. It is not clear how to optimally normalize the ratings from different domains to the same scale.

2) Product ranking single transfer: To handle the different rating scales, we can adopt product rankings from auxiliary domains as supervision, and enforce the relative rankings of the overlapping products to be the same as the transferred ranking. We first convert ratings (if any) from auxiliary domains to product rankings to eliminate difference in rating scales. After that, we let all auxiliary rankings be denoted by $\pi_1, \dots, \pi_K$, and define the indicator function on the pairs of products: $\mathbb{1}[v_1 \succeq_{\pi_k} v_2]$. To estimate a consensus product ranking out of the auxiliary rankings, define the following consensus score function: $\chi(v_1, v_2) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}[v_1 \succeq_{\pi_k} v_2]$. χ serves as a ranking agreement measurement, and the higher χ is, the more the auxiliary rankings agree upon the ordering of the two products. The weights of individual source are chosen to be uniform since without expert input, it can be hard to determine which source is more reliable than the other. In the experiments, we only retrieve the product pairs with their orderings agreed upon by all auxiliary rankings.

$$v_1 \succeq_{\bar{\pi}} v_2 \iff \chi(v_1, v_2) = 1 \tag{4}$$

Essentially, we have extracted a partial ordering of the overlapping products, denoted by $\bar{\pi}$. We incorporate $\bar{\pi}$ in the unsupervised bias correction procedure as follows. After calculating the product quality using Eq. (1), we solve the following optimization problem:

$$\begin{aligned} \mathbf{q}^* = \arg\min_{\mathbf{q}} \quad & \textstyle\sum_{j=1}^{n} \|q_j - \bar{q}_j\| \\ \text{s.t.} \quad & q_j \geq q_\ell \ \text{ if } \ v_j \succeq_{\bar{\pi}} v_\ell \end{aligned} \tag{5}$$

where $\bar{\mathbf{q}} = [\bar{q}_1, \dots, \bar{q}_n]$ is the product quality scores found by Eq. (1). This optimization problem models two objectives. First, the inferred scores of *all* products should be close to the averaged ratings obtained from the ratings, with reviewer bias taken into account. Second, the inferred scores of the overlapping products $q_i, i = 1, \dots, n$ are forced to follow the consensus ranking $\bar{\pi}$. Eq. (2) then takes the solution of the above optimization problem as input to estimate reviewer bias, and the iterations go on until convergence. The algorithm is called **CRST** (**C**onsensus **R**anking **S**ingle **T**ransfer).

B. A robust dual transfer approach

In product-centric single transfer approaches, the transferred rating/ranking is only used to correct possible rating/ranking bias in the target domain. We want to utilize the transferred rating/ranking in a more effective way by adding a "reviewer-centric" perspective and propose a novel dual transfer approach, which robustly applies the transferred knowledge to both products and reviewers. In the following, we first introduce the concept of "anchor reviewer reliability" that can facilitate reviewer-centric transfer learning.

1) Anchor reviewer reliability estimation and confident anchor reviewer identification: Anchor reviewers have reviewed both overlapping and non-overlapping products *in the target domain*, and serve as a bridge between the overlapping and non-overlapping products. If we can robustly estimate their reliability using their ratings/rankings on the overlapping products, and incorporate these reliability in bias correction, then one can expect a better ranking of the non-overlapping products. Assume that the transferred ranking of the overlapping products is of reasonable quality (though may not be perfect), the similarity between the transferred ranking and the product ranking provided by an anchor reviewer can be indicative of the reliability of the reviewer. We adopt Kendall-τ ranking correlation coefficient [8] to measure the reliability of anchor reviewers. Formally, reliability of an anchor reviewer is defined as

$$\tau(\pi_1, \pi_2) = \frac{2(C - D)}{n(n-1)} \tag{6}$$

where C (D, resp.) is the number of concordant (discordant, resp.) pairs of products in the two rankings π_1 (the transferred ranking) and π_2 (product ranking provided by the anchor reviewer). The function τ takes values in $[-1, 1]$ and a higher τ indicates the anchor reviewer is more reliable, and vice versa.

Note that if the denominator in Eq. (6) is small, then the estimated reliability of the anchor reviewer is less confident, thus we require there is a sufficiently large number of overlapping products between the transferred ranking and the product ranking provided by the anchor reviewer. However, if one requires a large number of overlapping products, too many anchor reviewers may be dropped off and bias correction may be affected. We will investigate this trade-off empirically later.

2) Incorporating reviewer reliability in the single transfer strategy: Now we need to use the anchor reviewers to help correct the bias. The bias of a reviewer is computed as in Eq. (2), but when computing the quality of a product (be it an overlapping one or not), we use the following equation:

$$q_j = \frac{1}{|\mathcal{R}_{\cdot j}|} \sum_{i \to j} r_{ij}(1 - b_i) \times (1 + rel(i)) \qquad (7)$$

The meaning of this equation is that, for an anchor reviewer, his/her rating for the j-th product should be amplified or discounted by his/her reliability, while for a regular reviewer, there is no effect of reliability and Eq. (7) is just the same as Eq. (1). The algorithm is called **CRDT** (**C**onsensus **R**anking **D**ual **T**ransfer) and summarized in Algorithm 1. Compared with **CRST**, **CRDT** has an additional step of anchor reviewer reliability estimation, and uses a different formula to estimate product scores, with anchor reviewer reliability taken into account. **CRDT** applies the transferred consensus ranking to both products (ranking constraint) and reviewers (reliability estimation) in the target domain.

Algorithm 1 Robust Bias Correction via Consensus Ranking Dual Transfer (**CRDT**)

```
Input: anchor reviewers $\{u_1, \ldots, u_s\}$, product ratings $\mathcal{R}$ in target domain, multiple external rankings $\pi_i, i = 1, \ldots, k$
Output: $q_j$ for the products.
Compute the consensus product ranking $\bar{\pi}$ from $\pi_i, i = 1, \ldots, K$, using Eq. (4).
for $i = 1 \to s$ do
    compute anchor reviewer reliability for $u_i$.
end for
while not convergent do
    Estimate reviewer bias using Eq. (2).
    Estimate unbiased product rating using Eq. (7).
    Enforce ranking of the overlapping products to agree with $\hat{\pi}$ by solving Eq. (5).
end while
```

C. Computational complexity analysis and incremental model update

We first consider the time complexity of building **CRDT** from scratches. The time complexity to compute a consensus ranking is linear in the number of ratings from all auxiliary sources. These computations can be distributed to multiple machines as there is no information sharing among sources. The space complexity $O(n)$ to store the averaged ratings of the products (instead of $O(n^2)$ to store the pairwise rank comparisons). Regarding calculating anchor reviewer reliabilities, a rather loose upper bound of the time complexity is $O(|\mathcal{R}|)$, namely the time complexity to go through all ratings. However, only a small portion of the reviewers are anchor reviewers, and only their ratings need to be visited during reliability calculation. The time complexity of estimating product quality and reviewer bias using Eq. (1) and (2) is $O(T * |\mathcal{R}|)$ wher T is the number of iterations needed for Algorithm 1 to converge. We show in the experiments that T is usually quite small and can be considered as a constant. Overall, both the time and space complexity of the proposed method is linear in $|\mathcal{R}|$.

Table II
CHARACTERISTICS OF RATING DATASETS

	NYC	PHX	SF
# of restaurants	79	77	85
# of users	9829	5804	8183
# of ratings	12415	8050	10186

Since the ratings keep accumulating, it is also important to consider incremental updates. It is trivial to update the product ratings for each auxiliary source. To update the reliabilities of the anchor reviewers, only their updated ratings will get involved, and that's a small number since normal reviewers don't usually add new ratings in a short period. Lastly, we only need to re-run Equations (1) and (2) once to update the solutions, using the q_i and b_j from previous iterations, since the bipartite graph and the reliabilities do not change significantly from previous iterations.

IV. EXPERIMENTS

A. Datasets and Performance Metrics

We employ a rating dataset collected from multiple rating websites as our test bed. Table II describes the rating data of three cities, New York City (NYC), Phoenix (PHX) and San Francisco (SF) from tripadvisor.com, which is our target domain. As external domain rankings, we collect ratings of the same set of restaurants from foursquare.com and yelp.com. Similar to [18], ratings from Zagat.com are treated as ground truths. We use the number of *concordant* pairs of products that are ordered consistently between the ground truths and the ranking derived by various bias correction algorithms. A good rating bias correction algorithm should produce more concordant pairs.

B. Baselines and experimental protocol

One can simply average the ratings of each product and derive a product ranking. We denote this method by "MEAN". This baseline does not take care of reviewer bias explicitly. We consider three more sophisticated baselines that explicitly consider rating bias. The unsupervised model proposed in [16] (denoted by "UN-SUP") iteratively and alternatively applies Eq. (1) and (2) until it converges. There are two baselines that exploit transferred knowledge, either by clipping the ratings of the overlapping products to the transferred ratings (called "S-SUP"), or by enforcing the ranking of the overlapping products to be the same as the transferred ranking (**CRST**). These two baselines do not consider and model reviewer reliability.

We randomly and uniformly select half of the products as overlapping products. The performance metric is computed on the non-overlapping parts, in order to check if the transfer knowledge can be propagated to the non-overlapping products and improve their ranking. We repeat the experiment 100 times and report the averaged performance. The proposed algorithm have a parameter to cut off the reviewers who have less confident estimation of their reliability. We fix this parameter to be 3 in the following results, and study the sensitivity of this parameter in Section IV-C1.

C. Results

Figure 1 compares the number of concordant pairs of products according to various rating debias methods on 3 datasets. From the figure, we can observe the followings. First, MEAN has significant lower performance than other methods among 2 out of 3 tasks (NYC and SF), and is slightly better than UN-SUP and CRST on the other task (PHX). One possible explanation of such mixed performance is that the average rating can sometimes remove the rating bias of individual reviewers, but can also fail to do so if a product is only rated by a few biased reviewers. Second, CRST and UNSUP have similar performance across all 3 datasets. This surprising fact indicates that the transferred ranking may be too difficult to be propagated to the non-overlapping products. The reason is that the transferred rankings are used to enforce the orderings of the overlapping products, while bias of reviewers is inferred indirectly, which is not very effective. Third, the performance of S-SUP is generally worse than that of UN-SUP and CRST, which is caused by the heterogeneity among different rating systems. Indeed, if the difference between rating scales is not handled carefully, simply normalizing and averaging ratings from different systems can be harmful. Lastly, we see that the proposed method performs the best. We conclude that the transferred rankings can indeed be used to find out the reliability of reviewers, which can effectively adjust the ratings on the non-overlapping products to more objective ratings, leading to a better ranking of the products.

1) Sensitivity study: An anchor reviewer has to rate more than a certain number of overlapping products to be qualified as a confident anchor reviewer. We set this threshold to be 2, 3 and 4, and report the performances of the proposed method, along with the performance of the best baseline, in Figure 2. We also indicate the number of anchor reviewers under each threshold on top of the bars in the figure. In general, we see that the proposed algorithm works best when the threshold is set to 2. The only setting our method is not as good as the best baseline is when the threshold is set to 3 on the NYC dataset. We have following conclusions. First, the performance of the proposed reviewer-centric approach is not that sensitive to the threshold. Even we set the threshold to as high as 5 and there are only tens of anchor reviewers, the CRDT method still outperforms the best baseline. Second, when the threshold is set to 2, and the reliability of an anchor reviewer is evaluated using the ordering of only 3 products. The resulting reliability estimation might not be very confident based on such a small sample. However, we can obtain a larger number of anchor reviewers to cover more non-overlapping products. The superior performance when the threshold is 2 indicates that the coverage of product by the anchor reviewers is more important than confidence of the reliability estimation. In practice, one can leave out a validation set to pick up the best threshold. The readers are referred to the full version[1] for more details, including the convergence of the proposed algorithm and its potential for suspicious reviewer detection.

V. Related Work

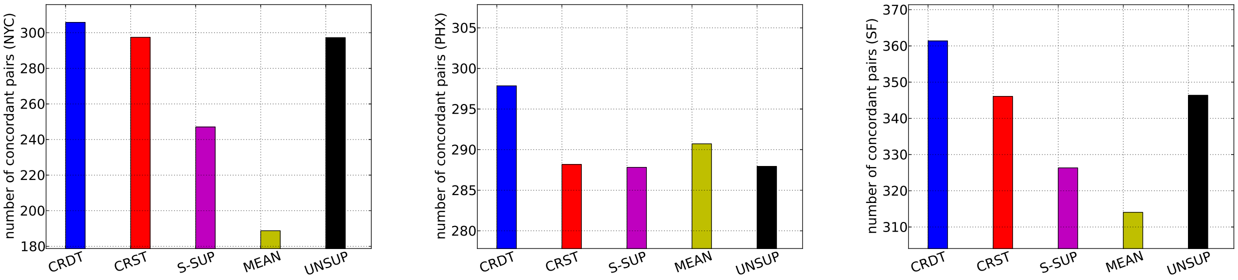
Correcting the inherent bias in ratings has attracted huge attentions. There are unsupervised [9, 10], semi-supervised and active learning approaches [16] towards this problem. Similar reinforcement approaches are adopted in [21, 20], but for different tasks. The approach proposed in this paper does not require any expert input, while still outperform previous unsupervised methods. In [5, 1, 17], they propose methods to detect fake or spamming reviews by looking at various features, such as user behaviors, review texts. Our method focuses on correcting the bias in the user-generated ratings, while the above works focus on text mining.

Transfer learning for ranking in the target domain can utilize more information to obtain a better ranking [7, 13]. More recently, they study several different strategies to combine ratings/rankings from multiple websites to estimate product quality [15]. We focus on addressing knowledge transfer for bias correction on non-overlapping products, while one may adopt any of the more sophisticated rank aggregation approaches. The proposed method utilize multiple related rankings to help bias correction, where multiple rankings are “averaged” for knowledge transfer. Such ranking averaging has been studied intensively for a long history [3, 2, 6, 4, 11]. The dual transfer learning methodology is first proposed in [12], which studies the strategy in the context of classification, while this work proposes a novel dual transfer strategy for the rating bias correction problem.

VI. Conclusions

In this paper we study the problem of how to transfer knowledge from multiple external rankings to improve the ranking of non-overlapping products in the target domain, where the external rankings only partially overlap with the target domain. We propose a framework to reliably extract, transfer and utilize transferable knowledge to radically improve the target ranking task. Alternative knowledge transfer methods are also explored and we find that a dual transfer approach works best among all strategies, including existing unsupervised and semi-supervised methods. Experimental results show that the proposed method delivers promising improvement, and the potential to use the inferred reviewers' bias as anomaly detection signals.

[1] http://www.cs.uic.edu/~sxie/papers.html

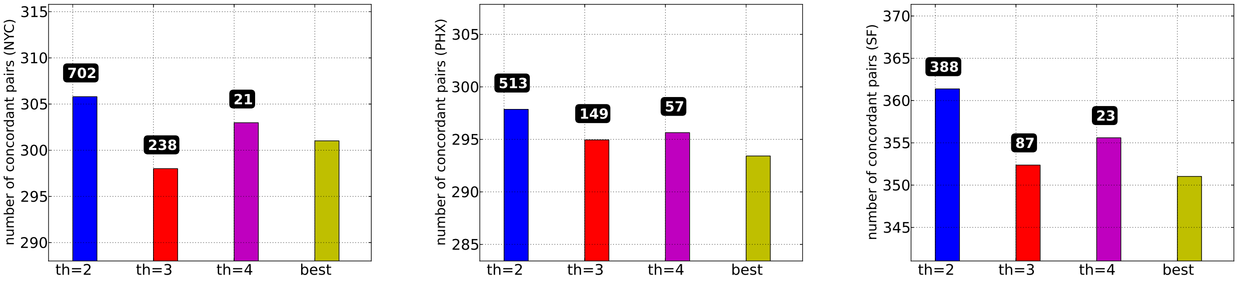


(a) # concordant pairs (NYC)

(b) # concordant pairs (PHX)

(c) # concordant pairs (SF)

Figure 1. Overall comparisons of the proposed method and the baselines



(a) NYC

(b) PHX

(c) SF

Figure 2. Sensitivity of CRDT (“best” indicates the baseline with the best performance)

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] Mukherjee A, Liu B, and Glance N. Spotting fake reviewer groups in consumer reviews. WWW ’12.

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

[5] Lim E-P, Nguyen V-A, Jindal N, Liu B, and Lauw H W. Detecting product review spammers using rating behaviors. CIKM ’10.

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

[7] Qingbo Hu, Guan Wang, and P.S. Yu. Transferring influence: Supervised learning for efficient influence maximization across networks. In *CollaborateCom*, 2014.

[8] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 1938.

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

[10] Hady W. Lauw, Ee-Peng Lim, and Ke Wang. Summarizing review scores of unequal reviewers. SDM, 2007.

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

[12] Mingsheng Long, Jianmin Wang, Guiguang Ding, Wei Cheng, Xiang Zhang, and Wei Wang. Dual transfer learning. SDM, 2012.

[13] Chun-Ta Lu, Sihong Xie, Xiangnan Kong, and Philip S. Yu. Inferring the impacts of social media on crowdfunding. In *WSDM*, 2014.

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

[15] Mary McGlohon, Natalie S. Glance, and Zach Reiter. Star quality: Aggregating reviews to rank products and merchants. In *ICWSM*, 2010.

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

[17] Jindal N and Liu B. Opinion spam and analysis. WSDM ’08.

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

[19] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. reCAPTCHA: Human-based character recognition via web security measures. 2008.

[20] 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.

[21] 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.