

Singapore Management University Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

11-2008

Bias and Controversy in Evaluation Systems

Hady Wirawan LAUW

Singapore Management University, hadywlaw@smu.edu.sg

Ee Peng LIM


Singapore Management University, eplim@smu.edu.sg

Ke WANG

Simon Fraser University

DOI: <https://doi.org/10.1109/tkde.2008.77>

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

Citation

LAUW, Hady Wirawan; LIM, Ee Peng; and WANG, Ke. Bias and Controversy in Evaluation Systems. (2008). *IEEE Transactions on Knowledge and Data Engineering*. 20, (11), 1490-1504. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/127

This Journal Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Bias and Controversy: Beyond the Statistical Deviation

Hady W. Lauw
Nanyang Technological
University

hadylauw@pmail.ntu.edu.sg

Ee-Peng Lim
Nanyang Technological
University

aseplim@ntu.edu.sg

Ke Wang^{*}
Simon Fraser University
wangk@cs.sfu.ca

ABSTRACT

In this paper, we investigate how deviation in evaluation activities may reveal bias on the part of reviewers and controversy on the part of evaluated objects. We focus on a ‘data-centric approach’ where the evaluation data is assumed to represent the ‘ground truth’. The standard statistical approaches take evaluation and deviation at face value. We argue that attention should be paid to the subjectivity of evaluation, judging the evaluation score not just on ‘what is being said’ (deviation), but also on ‘who says it’ (reviewer) as well as on ‘whom it is said about’ (object). Furthermore, we observe that bias and controversy are mutually dependent, as there is more bias if there is higher deviation on a less controversial object. To address this mutual dependency, we propose a reinforcement model to identify bias and controversy. We test our model on real-life data to verify its applicability.

Categories and Subject Descriptors: H.4 [Information Systems Applications]; J.4 [Social and Behavioral Sciences]

General Terms: Algorithms, Experimentation, Measurement

Keywords: bias, controversy, evaluation, social network

1. INTRODUCTION

Evaluation or assessment is a fundamental activity in our life because it touches on various areas of human concerns. Students evaluate instructors; referees evaluate athletes; reviewers evaluate submitted papers. Online evaluation is just as prevalent, if not more. For example, product review sites allow users to assign ratings to goods, such as www.amazon.com and www.imdb.com. In any evaluation, the key questions include whether the evaluation is “fair”, whether reviewers are “biased”, whether a large deviation is normal. For instance, an article in “The Scientist” [12] raises such questions on the peer review practice. Another

^{*}The third author’s work was done while he was visiting Nanyang Technological University, Singapore.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD’06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

Copyright 2006 ACM 1-59593-339-5/06/0008 ...\$5.00.

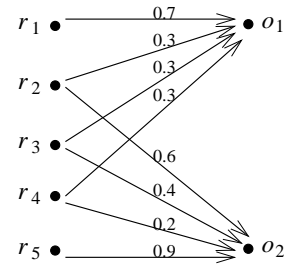


Figure 1: Example Scenario (with e_{ij} values)

example is *buzz or stealth marketing* [15] where companies hire people to post the ratings of products. The questions on “bias” and “fairness” are important and yet are very difficult to define and answer. For one reason, “bias” and “fairness” are subjective in that different people have different views. For another reason, reviewers and evaluated objects are not “uniform”. Some objects are more controversial and a larger deviation among reviewers is expected.

In this paper, we study the notions of “bias” of reviewers and “controversy” of objects evaluated. Given the subjective nature of “bias” and “controversy”, a complete answer to these questions goes way beyond computer science. Therefore, our study focuses on a “data-centric approach”, where “bias” and “controversy” can be objectively quantified from the evaluation scores given by reviewers to objects. In this approach, we assume that the evaluation scores represent the “ground truth” that can be trusted for the study. In particular, there is no fraud in the assignment of reviewers to an object. For example, if the reviewers for an object are chosen deliberately in favor of or against the object and if all reviewers give similar scores to the object, the “data-centric approach” is not able to identify the “bias” caused by such systematic frauds. Essentially, our approach assumes that most reviewers are “honest” in the sense of acting according to their best judgment and yet can still be “biased”.

Moreover, we do not try to identify the causes of “bias”. There are too many possible reasons behind “bias” and reviewers may have been influenced by different ones. Instead, we focus on identifying and measuring the manifestation of “bias”, which in turn can be used to investigate the causes of “bias”. The same can be said for “controversy”. We formally define our notions of “bias” and “controversy” shortly.

In its most basic construct, an evaluation system consists of the type of reviewers and the type of objects. A reviewer

r_i may assign to an object o_j an evaluation score $e_{ij} \in [0, 1]$. Here, we use the terms *reviewer* and *object* in the general sense, referring not to what they are, but to their respective roles. It may well be the case that both reviewers and objects are of the same type (e.g., person). A bipartite graph representation is given in Figure 1. The values given in the figure are e_{ij} values, e.g., $e_{11} = 0.7$. For each e_{ij} , we may derive $d_{ij} \in [0, 1]$, which measures r_i 's deviation from the consensus (such as mean or median) of o_j . Given such a graph, we seek to measure the bias value $b_i \in [0, 1]$ of each r_i and the controversy value $c_j \in [0, 1]$ of each o_j .

A straightforward solution to the problem stated above is to employ standard statistical measures. Assuming that d_{ij} is known, the bias value b_i may simply be the average deviation by r_i on all objects she has evaluated, as given in Equation 1. The controversy value c_j may simply be the average deviation on o_j by all reviewers evaluating it, as given in Equation 2. We call this pair of equations the *Naive* solution.

$$b_i = \text{Avg}_j d_{ij} \quad (1)$$

$$c_j = \text{Avg}_i d_{ij} \quad (2)$$

For example, for the scenario in Figure 1, the *Naive* solution would conclude that r_1 is less biased than r_5 . Here, we derive d_{ij} as the absolute distance from e_{ij} to the mean evaluation by all reviewers of o_j . For instance, we have $d_{11} = 0.3$ and $d_{52} = 0.375$. Since r_1 evaluates only o_1 and r_5 evaluates only o_2 , according to Equation 1, $b_1 = d_{11} = 0.3$ and $b_5 = d_{52} = 0.375$. We see that $b_1 < b_5$, concluding r_1 is less biased than r_5 .

1.1 Bias and Controversy

The *Naive* approach is akin to taking deviation at its face value. It is therefore naive as it treats all reviewers and objects equally. To use an analogy, the approach is to take into account only ‘what is being said’ (deviation) while ignoring ‘who says it’ (reviewer) and ‘about whom it is said’ (object). However, deviation could have arisen due to either bias or controversy. Thus, there is a need to pay attention to the particular reviewer or object that a deviation concerns.

It is further observed that bias and controversy are inter-related quantities. When determining how biased a reviewer is, we should use deviation attributed to the bias of this reviewer, and not to the controversy of evaluated objects. Similarly, when determining how controversial an object is, we should use deviation attributed to the controversy of this object, and not to the bias of evaluating reviewers.

In this paper, we investigate the two main issues ignored by the *Naive* model in quantifying bias and controversy.

Subjectivity In determining bias and controversy, we should look beyond deviation. Here, *subjective* does not mean that the outcome of analysis is different to different people. Rather, we refer to the *objective subjectivity* of how deviation should be seen in the context of the concerned reviewer or object, as given by the data.

Mutual Dependency Bias and controversy are mutually dependent upon each other. Determining the bias of a reviewer requires knowing the controversy of objects she has evaluated, and vice versa.

We now present the following observation of bias and controversy that underlines our basic approach to this problem.

Bias *A reviewer is more biased if there is more deviation on a less controversial object.*

Controversy *An object is more controversial if there is more deviation by a less biased reviewer.*

Re-examining the example in Figure 1, based on the above observation, we now argue that r_1 is in fact more biased than r_5 . Visual inspection would reveal that co-reviewers of o_1 are much more in agreement (with 3 out of 4 reviewers agreeing on the score) than co-reviewers of o_2 (with 4 divergent scores). Therefore, o_2 is more controversial than o_1 because there is a lack of consensus among o_2 's reviewers. However, in this case, r_5 may not be biased as deviation d_{52} may be attributed to the controversy of o_2 . On the other hand, r_1 deviates on an object that her co-reviewers could agree upon, implying bias on her part. Thus, *Naive* has incorrectly concluded that r_1 is less biased than r_5 .

1.2 Contributions

We present a new approach to the problem of quantifying bias and controversy within an evaluation system. First, we propose the above observation that incorporates a new notion of mutual dependency between bias and controversy. This will subsequently be developed into a reinforcement-based model. Interestingly, this model has an underlying presumption resulting in the so-called ‘no evidence cases’. Moreover, we also examine several issues that significantly affect the outcome of this model. Finally, we conduct experiments on real-life data to analyze how our proposed approach is different from the *Naive* approach.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 develops the framework of our reinforcement model. Section 4 highlights several issues that would influence the outcome of the model. Section 5 seeks to verify the model's applicability through experiments on a real dataset. Section 6 concludes this paper.

2. RELATED WORK

The study of evaluation systems has also been conducted in fields outside computer science. In management science, [5] looks at how in evaluating start-up teams, venture capitalists seem to favor those similar to themselves, while [13] investigates how using a diversity of objective and subjective measures in performance evaluation may affect the evaluation scores assigned. Different from these works, we do not factor in any other information besides the evaluation scores assigned by reviewers to objects.

The bipartite structure of evaluation systems resembles two-mode social networks (consisting of two types of nodes). One classic problem in such networks is identifying the central nodes [4] or those playing mediative role in facilitating linkages among nodes of both types. Alternatively, anomalous nodes [16] or those with low affiliation to any neighborhood may be of interest.

Beyond bipartite structures, there are also other works on social networks or Web graphs, such as identifying the most influential individuals [8] or finding the most interesting connections among several nodes [3] or grouping together related Web pages into Web communities [6]. None of these works is concerned with evaluation data. Moreover, most

deal with non-directional relationships and graph topologies. Thus far, we have not come across any existing work that addresses the issues of bias and controversy.

The iterative computation method used to implement the mutual dependency in this paper has first been addressed in linear algebra [1, 7]. Several other works have also made use of mutual dependency property, notably as applied to ranking pages for Web search [9, 14] and ranking products based on propagated profitability [17]. However, such works are primarily based on the notion of popularity, which is not congruent with either bias or controversy. Simply evaluating the most number of objects would not imply bias.

Finally, our work also has some relation to outlier detection [10], which is concerned with identifying points, from a set of points, that are far away from the majority of points. In a way, the distance measure is similar to the statistical approach of taking deviation at face value. For instance, high deviation by a biased reviewer may mark her as an outlier. However, in our approach, how outlying such a reviewer is would depend on how controversial the concerned object is. These notions of subjectivity and mutual dependency again are not usually factored into outlier detection problems.

3. INVERSE REINFORCEMENT MODEL

In this section, we develop a computational model of bias and controversy that factors in their mutual dependency. While the raw data would likely contain evaluation score e_{ij} , as a matter of generality, in our model development, we work with deviation value d_{ij} . More on how deviation may be derived from evaluation will be discussed in Section 4.1.

The *Naive* model simply translates more deviation to more bias. In fact, deviation could have arisen because of either bias *or* controversy. As much as possible, we should attribute to a reviewer only deviation due to her bias, and not to the controversy of evaluated object. Therein lies our proposed approach: to reduce the amount of deviation attributed to bias by the amount of controversy that could have contributed to that deviation.

This approach is summed up by the pair of Equation 3, which determines bias, and Equation 4, which determines controversy. Here, we use \bar{b}_i and \bar{c}_j to denote the complements of b_i and c_j respectively, which means that \bar{b}_i and \bar{c}_j grow inversely with b_i and c_j respectively. Moreover, we use *Agg* to represent the class of aggregate functions to combine the relevant values over i or j respectively. An appropriate aggregate function should yield a value that is representative of a reviewer or an object’s “behavior”. We avoid using *summation* so as not to incorporate the notion of popularity. Possible options include *minimum*, *maximum*, and *average*. Particularly, *average* is an intuitive choice as it takes into account repeated deviation by a biased reviewer.

$$b_i = \underset{j}{\text{Agg}} d_{ij} \cdot \bar{c}_j \quad (3)$$

$$c_j = \underset{i}{\text{Agg}} d_{ij} \cdot \bar{b}_i \quad (4)$$

The above equations reflect the inversely proportional relationship between bias and controversy, which gives this model its name: *Inverse Reinforcement* or *IR* model. A reviewer’s bias value is higher for high deviation on objects with low controversy (high \bar{c}_j values). An object’s contro-

versy value is higher for high deviation by reviewers with low bias (high \bar{b}_i values).

A reviewer’s deviation values on less controversial objects would better reveal her bias, as more controversial objects reflect their own controversy. In a way, we rely more on the less controversial objects (high \bar{c}_j values) as “evidence” to reveal bias. However, in the case where a reviewer evaluates only very controversial objects, there is no “evidence” to reveal her bias. Thus, we refer to reviewers who have evaluated only very controversial objects as “no evidence cases”. Such “no evidence cases” will be assigned low bias values by the *IR* model (Equation 3). Similar remarks can be made on the controversy of objects. Hence, the *IR* model adopts the presumption below.

Presumption

- *A reviewer is presumed not biased unless proven biased.*
- *An object is presumed not controversial unless proven controversial.*

Therefore, in the presence of “no evidence cases”, this presumption bears the following implications on bias (similarly on controversy).

1. We have more “confidence” on those assigned high bias values, as they would have come about due to high deviation and low controversy.
2. We have less “confidence” on those assigned low bias values, as some of them may have evaluated only controversial objects (“no evidence cases”).

Because the data may potentially contain “no evidence cases”, we focus on the more “confident” aspects of the outcome of *IR* model. *We recommend that IR should primarily be used for identifying biased reviewers and controversial objects.* However, in practice, several steps may be taken to avoid “no evidence cases”, such as (1) having reviewers review more objects to increase the probability of having non-controversial objects or (2) ensuring each reviewer is allocated at least a few non-controversial objects, assuming we have prior knowledge about the controversy of objects.

Note also that when the effects of mutual dependency are removed, *IR* degenerates into *Naive*. For instance, by fixing \bar{c}_j as a constant in Equation 3, the b_i values determined by Equation 3 will have the same ordering as those determined by Equation 1.

4. OTHER ISSUES

Several issues that may affect the effectiveness of *IR*’s application include the derivation of deviation from evaluation and the convergence of the iterative method for *IR*.

4.1 Deviation

In Section 1, we introduce one way to derive deviation from evaluation, which we now term *deviation from mean*. It takes deviation d_{ij} as the absolute distance between r_i ’s evaluation and the mean evaluation by all reviewers r_k of object o_j . Equation 5 gives the definition of this measure, where m_j denotes the number of reviewers of o_j .

$$d_{ij} = |e_{ij} - \frac{1}{m_j} \sum_{r_k} e_{kj}| \quad (5)$$

Another possible deviation measure is *deviation from co-reviewers*. This measure takes deviation d_{ij} as the average distance between r_i 's evaluation and the evaluation by each co-reviewer r_k . For $m_j > 1$ number of reviewers of o_j (including r_i), d_{ij} can be worked out according to Equation 6.

$$d_{ij} = \frac{1}{m_j - 1} \sum_{r_k \neq r_i} |e_{kj} - e_{ij}| \quad (6)$$

For example, suppose reviewers r_1 , r_2 , and r_3 assign the following evaluations $e_{1j} = 0.0$, $e_{2j} = 0.5$, and $e_{3j} = 1.0$ on the same object o_j . For this case, *deviation from mean* would yield the following deviation values, $d_{1j} = 0.5$, $d_{2j} = 0.0$, $d_{3j} = 0.5$, claiming that r_2 has not deviated at all. On the other hand, *deviation from co-reviewers* would yield $d_{1j} = 0.75$, $d_{2j} = 0.50$, $d_{3j} = 0.75$. Firstly, *deviation from mean*'s claim that r_2 has not deviated at all is not reasonable, as clearly all the reviewers do not agree on o_j 's evaluation score. Furthermore, we think that *deviation from co-reviewers*'s claim that d_{1j} is 1.5 times d_{2j} ($0.75 \div 0.50$) is more reasonable than the *deviation from mean*'s claim that d_{1j} is infinitely greater than d_{2j} ($0.50 \div 0$).

The above example highlights the weakness of *deviation from mean*, which is more likely to produce deviation values close to zero. As the number of reviewers of an object grows, the distribution of evaluation scores would likely peak at or near the mean. Deviation from the mean would then approach zero. This is disadvantageous because the ratio among deviation values would determine the outcome of computation. Very small values mean that a small change in absolute value may trigger a large change in ratio, making the system potentially too sensitive to small changes. Hence, *deviation from co-reviewers* is our recommended measure as it is more likely to have a distribution of d_{ij} away from zero. We also use this deviation measure in the implementation of the *Naive* and *IR* models for experiments.

4.2 Convergence

The computation of bias and controversy in *IR* can be modeled as a problem of finding an eigenvector of a square matrix. *Average* is the aggregate function used for the following computation. We also assume that the linear relationships $b_i + \bar{b}_i = 1$ and $c_j + \bar{c}_j = 1$ hold¹. We may then re-write Equations 3 and 4 as Equations 7 and 8 respectively. n denotes total number of objects; m denotes total number of reviewers; n_i denotes number of objects evaluated by r_i ; and m_j denotes number of reviewers evaluating o_j .

$$b_i = \underset{j}{Avg} d_{ij} \cdot (1 - c_j) = \frac{\sum_{j=1}^n d_{ij} \cdot (1 - c_j)}{n_i} \quad (7)$$

$$c_j = \underset{i}{Avg} d_{ij} \cdot (1 - b_i) = \frac{\sum_{i=1}^m d_{ij} \cdot (1 - b_i)}{m_j} \quad (8)$$

Our matrix representation for *IR* is then as follows. We represent the $m \times 1$ vector of b_i values as B , $n \times 1$ vector of c_j values as C , column vector of appropriate length whose all elements are all 1's as $\mathbf{1}$, and $m \times n$ matrix of d_{ij} as D . From D , we may derive two other matrices, I whose each element

¹There are other options of defining complement mathematically, such as making \bar{b}_i the reciprocal of b_i , but such options are not explored in this paper.

is $d_{ij} \div n_i$ for corresponding i , and J whose each element is $d_{ij} \div m_j$ for corresponding j . Then Equations 7 and 8 can be re-written as matrix Equations 9 and 10 respectively.

$$B = I (\mathbf{1} - C) \quad (9)$$

$$C = J^T (\mathbf{1} - B) \quad (10)$$

By substituting Equations 9 and 10 into each other, we have recursive Equations 11 and 12.

$$B = (I\mathbf{1} - IJ^T\mathbf{1}) + IJ^TB \quad (11)$$

$$C = (J^T\mathbf{1} - J^TI\mathbf{1}) + J^TIC \quad (12)$$

Suppose for any $w \times 1$ column vector W , we use the notation W^m to denote $w \times m$ matrix formed by replicating W across m columns. If B is L_1 -normalized, i.e., $\sum_{i=1}^m |b_i| = 1$, then $W^m B = W$ holds. We use this notation to transform the previous equations into equivalent Equations 13 and 14.

$$B = (I\mathbf{1} - IJ^T\mathbf{1})^m B + IJ^TB \quad (13)$$

$$C = (J^T\mathbf{1} - J^TI\mathbf{1})^n C + J^TIC \quad (14)$$

Factorizing out B from the right-hand side of Equation 13 and C from the right-hand side of Equation 14 would yield recursive forms $B = XB$ and $C = YC$. The iterative process for B is given by $B_{k+1} = XB_k$, where the output of the k -th iteration is used as input for the $(k+1)$ -th iteration. Subject to the assumption that the square asymmetric matrix X is *diagonalizable* (it has linearly-independent eigenvectors) and has a uniquely largest eigenvalue [7], then as k increases, B_k will converge to the dominant eigenvector of X almost independently of the initial B_0 . If desired, these conditions for convergence can be tested, for instance by inspecting the eigenvalues or eigenvectors of the square matrix [1]. In that case, eigenvalues or eigenvectors may be determined using other methods such as [11]. Experimentally, convergence can be observed as stable B values (after normalization) across consecutive iterations. Convergence for C can be similarly argued.

Normalization Before each iteration, the input vector (e.g., B) is normalized. Normalization maintains an invariant state between two consecutive iterations so that convergence can be observed as no change or very little change in values. It involves dividing elements of a vector by a constant, such that their relative ratio remains unchanged.

L_p normalization of a vector B results in $\sum_{i=1}^m |b_i|^p = 1$. Commonly L_1 or L_2 is used [9, 14]. The summation means that as m increases, the individual b_i approaches zero. When $b_i \rightarrow 0$, *IR* (Equation 8) may degenerate into *Naive* (Equation 2). To counter the effect of summation, higher values of p could be used. We employ L_∞ normalization, which is equivalent to dividing vector elements by the largest one. The largest element after L_∞ normalization is 1.

5. EXPERIMENTS

The objective is to compare the efficacy of *Naive* and *IR* in identifying bias and controversy. First, exemplary reviewers

	d_{ij}	Rank	
		Naive	IR
user-dlockeretz		16	4
mvie_mu-1059489	0.400	78	90
mvie_mu-1032176	0.200	84	93
mvie_mu-1028572	0.029	92	92

Table 1: Bias Rank for *user-dlockeretz*

and objects are examined. Then, ranked lists by *Naive* and *IR* are compared using various similarity measures.

All our experimental runs involve few iterations and converge in less than a second. Computational complexity is not an important issue and will not be further examined.

5.1 Data

The data is acquired by crawling the product review Web site Epinions (www.epinions.com) for two days, starting with the seed page “Epinions Top Reviewers in Books”². The crawled Web pages represent a subset of all products, reviewers and evaluation ratings available from Epinions. The subset consists of 57320 web pages capturing 3797 products, 14607 reviewers, and 24008 evaluation ratings.

We impose several filtering conditions to make the data more suitable for experiments. Firstly, we prune the network such that all products have at least 5 reviewers and all reviewers have rated at least 3 products. This weeds out the occasional reviewers and products and gives greater support in determining a reviewer or product’s “behavior”. Any higher threshold would result in too small a network. Epinions assigns each product a category. The three most popular categories in the dataset are *books*, *videos*, and *music*. After filtering, only *videos* has a significant network size, with 113 products (objects), 138 reviewers, and 910 evaluation ratings. This category is selected for further analysis.

Since the focus of the experiments is not on scalability, the selection of data is not so much driven by the size of the data. The data selected is reasonably large for the propagation effect to take place within the network, and yet is not so overly large that analysis of the results is made difficult.

A reviewer assigns 0 to 5 stars to an object, with 5 being the best. We rescale these evaluation scores to a range from 0 to 1 by a simple division by 5 (e.g., 2 stars is 0.4).

5.2 Case Examples

Below are specific examples contrasting how *Naive* and *IR* determine biased reviewers and controversial objects.

Biased (IR) vs. Less Biased (Naive) Reviewers are placed in ranked lists in descending order of bias values (highest bias value is rank 1) as computed by *Naive* and *IR* respectively. First, we look at a reviewer who is assigned a lower bias rank by *Naive* than by *IR*. *user-dlockeretz*, whose profile is given in Table 1, is ranked 16 by *Naive* and 4 by *IR*. This profile includes d_{ij} values and controversy ranks (highest controversy value is rank 1) of the objects she reviewed. Notably, *user-dlockeretz* has high deviation on the first two objects. These objects also have very low controversy ranks by *IR* (ranks 90 and 93 out of 113). Furthermore, these

²http://www.epinions.com/member/community_lists.html/show/~6/display_list/~true/vert/~3321654/year/~1900/sec/~community_member_list/pp/~1/pa/~1

	d_{ij}	Rank	
		Naive	IR
mvie_mu-1016922		13	8
user-milymac	0.400	31	41
user-susidee34	0.320	71	98
user-moonmoods_52	0.240	99	97
user-pmills1210	0.160	39	45
user-andrew_hicks	0.160	62	67
user-mfunk75	0.160	110	108

Table 2: Controversy Rank for *mvie_mu-1016922*

	d_{ij}	Rank	
		Naive	IR
user-ynmaeven		10	54
mvie_mu-1019525	0.375	12	9
mvie_mu-1023730	0.267	2	2
mvie_mu-1084155	0.200	7	5
mvie_mu-1041911	0.160	14	22
mvie_mu-1102698	0.127	31	25

Table 3: Bias Rank for *user-ynmaeven*

two objects are given lower controversy ranks by *IR* than by *Naive*. Given these objects’ low controversy, *IR* takes the high d_{ij} values more seriously. *Naive* ignores these objects’ low controversy and decides based on deviation alone.

Controversial (IR) vs. Less Controversial (Naive)

Next, we examine an object given a lower controversy rank by *Naive* than by *IR*. For instance, *mvie_mu-1016922* (Table 2), is ranked 13 by *Naive*, but 8 by *IR*. Looking at the bias ranks of *mvie_mu-1016922*’s reviewers, we see that some of these reviewers have very low bias ranks (ranks 97, 98, 108 out of 138). Also, the bias ranks assigned by *IR* to *mvie_mu-1016922*’s reviewers tend to be lower than those by *Naive*. Deviation by reviewers with low bias values would better reflect *mvie_mu-1016922*’s controversy. *Naive* ignores this notion of subjectivity, and decides on a lower controversy rank of *mvie_mu-1016922* based solely on deviation values.

Less Biased (IR) vs. Biased (Naive) We present one example where *Naive*’s claim of bias is not really substantiated. Consider *user-ynmaeven* whose profile is given in Table 3. The objects on which *user-ynmaeven* has highest deviation on also have very high controversy ranks (ranks 2, 5, 9). The high deviation values could be attributed to the high controversy of these objects. There is no substantial case to claim that *user-ynmaeven* is really biased.

Due to space constraint, example of the only other case (objects determined to be highly controversial by *Naive* but less by *IR*) is not given here, but similar results apply.

5.3 Comparison of Ranked Lists

To see if the differences between *IR* and *Naive* surface on a larger scale as well, we compare greater subsections of the ranked lists (top 10%, 20%, 30%). We focus on the most biased (or controversial) ends of the ranked lists as these are the ends targeted by *IR* (see Section 3).

5.3.1 Measures of Comparison

For comparing two ranked lists, we use three similarity functions originally proposed to compare various permutations [2], with some adaptations for our needs.

	reviewers			objects		
	10%	20%	30%	10%	20%	30%
<i>Overlap</i>	0.64	0.71	0.76	0.73	0.78	0.88
<i>Kendall</i>	0.75	0.70	0.77	0.84	0.84	0.82
<i>Spearman</i>	0.67	0.60	0.68	0.73	0.73	0.72

Table 4: Most Biased and Controversial: *Naive* vs. *IR*

Overlap Similarity between two ranked lists is the proportion of items common to both lists. For two ranked lists τ_1 and τ_2 of length n , where A is the set of items in τ_1 and B the set of items in τ_2 , the *Overlap* similarity between the two lists can be evaluated as shown in Equation 15. *Overlap* similarity ranges from 0 (disjoint) to 1 (total overlap).

$$Overlap(\tau_1, \tau_2) = \frac{|A \cap B|}{n} \quad (15)$$

Kendall Similarity [2] counts the number of pairs for which the two ranked lists agree on their ordering. Hence, *Kendall* similarity penalizes for each pair of items (x, y) where $rank(x) > rank(y)$ in one list but $rank(x) < rank(y)$ in the other list. Equation 16 shows how this similarity is evaluated. *Kendall* similarity ranges from 0 (completely reversed) to 1 (completely identical).

$$Kendall(\tau_1, \tau_2) = \frac{|\{(x, y) \mid \tau_1 \text{ and } \tau_2 \text{ agree on order of } (x, y)\}|}{\frac{1}{2}n(n-1)} \quad (16)$$

Spearman Similarity [2] counts, for each item x , the difference between its rank in the first list $rank_1(x)$ and its rank in the second list $rank_2(x)$. The aggregate differences across all items in the list contribute to the final similarity score as in Equation 17. The normalization denominator is $N = n^2/2$ for even values of n and $N = (n+1)(n-1)/2$ for odd values of n . *Spearman* similarity ranges from 0 (completely reversed) to 1 (completely identical).

$$Spearman(\tau_1, \tau_2) = 1 - \frac{\sum_{x=1}^n |rank_1(x) - rank_2(x)|}{N} \quad (17)$$

5.3.2 Comparison Results

The top $k\%$ of two lists may not contain the same set of reviewers or objects. For *Kendall* and *Spearman*, we take the top $k\%$ items of *IR* as the reference set. We then construct a ranked list of the same items according to their ordering in *Naive*, with any gap between rank orders removed.

As Table 4 shows, similarity values between ranked lists produced by *Naive* and *IR* range from 0.60 to 0.88. We do not expect the ranked lists by *Naive* and *IR* to be completely different. This is because in any typical evaluation system most reviewers (and objects) would “behave normally”. Nevertheless, the similarity values in Table 4 suggest that there are significant differences between the ranked lists by *Naive* and *IR*. These differences come about due to cases, such as those in Section 5.2, where *IR* disagrees with *Naive*. Unlike *Naive*, *IR* takes into account the mutual dependency between bias and controversy. We further observe that *Spearman* values are generally the lowest compared to the other similarity values. *Spearman* compares exact ranks, which implies that even if *Naive* and *IR* may feature the same reviewers/objects, their ranks in respective lists would be different.

6. CONCLUSION

In this paper, we seek to quantify the notions of bias and controversy within an evaluation system. Deviation is a common occurrence in evaluation activities, and significant deviation may help reveal bias of reviewers or controversy of objects. However, statistical measures tend towards objectivity, taking deviation values as they are. Here, we propose tackling the problem based on two major issues: (1) *subjectivity*, taking into account bias of reviewer and controversy of object related to deviation and (2) *mutual dependency*, recognizing that quantifying bias requires knowing controversy and vice versa. We have proposed the *Inverse Reinforcement* or *IR* model based on these ideas. Another contribution is in working out several crucial issues that might affect the outcome, such as derivation of deviation as well as convergence of iterative computation of *IR*. We have also sought to verify the proposed model through experiments with real-life data, and the results have been encouraging.

7. REFERENCES

- [1] H. Anton and C. Rorres. *Elementary Linear Algebra with Applications*. John Wiley & Sons, Inc., 1987.
- [2] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.
- [3] C. Faloutsos, K. S. McCurley, and A. Tomkins. Fast discovery of connection subgraphs. In *KDD*, pages 118–127, 2004.
- [4] K. Faust. Centrality in affiliation networks. *Social Networks*, 19(2):157–191, 1997.
- [5] N. Franke, M. Gruber, D. Harhoff, and J. Henkel. What you are is what you like – similarity biases in venture capitalists’ evaluations of start-up teams. *Journal of Business Venturing*, In Press, Corrected Proof, 2005.
- [6] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web communities from link topology. In *Hypertext*, pages 225–234, 1998.
- [7] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.
- [8] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.
- [9] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [10] E. Knorr and R. Ng. A unified notion of outliers: Properties and computation. In *KDD*, pages 219–222, 1997.
- [11] Mathematica. Mathematica. <http://www.wolfram.com/products/mathematica/index.html>.
- [12] A. McCook. Is peer review broken? *The Scientist*, 20(2):26, 2006.
- [13] F. Moers. Discretion and bias in performance evaluation: The impact of diversity and subjectivity. *Accounting, Organizations and Society*, 30(1):67–80, 2005.
- [14] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. In *Stanford Digital Library Technologies Project*, 1998.
- [15] G. Ruskin. Commercial Alert asks FTC to investigate buzz marketers for deception. Retrieved February 17, 2006, from <http://www.commercialalert.org/news/news-releases/2005/10/commercial-alert-asks-ftc-to-investigate-buzz-marketers-for-deception>, 2005.
- [16] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM*, pages 418–425, 2005.
- [17] K. Wang and M.-Y. T. Su. Item selection by “hub-authority” profit ranking. In *KDD*, pages 652–657, 2002.