Creating Incentives for Honest Rating Submission by Rating the Raters

Jörg H. Lepler
Cambridge University
Computer Laboratory Faculty of Economics
Joerg.Lepler - at - cl.cam.ac.uk

Abstract

Reputations of electronic parties are important for their economic viability in business settings. Potential clients of e-services cannot rely solely on their own experiences to gain knowledge about e-services' reputations. Plainly aggregating rating comments from other clients can often be misleading. Therefore, we are presenting a reputation aggregation model which seeks for the subgroup of raters that (1) contains the largest degree of overall agreement and (2) derives the resulting reputation scores from their comments. We complement the reputation system by adding policies that promote both truth revelation for the rater commentary, and also competition over service quality between providers. Then we describe the recursive algorithm of the model, which judges a rater's commenting quality based on the divergence between his and the other raters' comments. This algorithm feeds the commenting quality back into the aggregation function as a weight on the rater's comments. We evaluate our algorithm in simulations of three challenging threat scenarios. To this end, we show that our aggregation model can be effectively used to deselect weak rating accuracy, and to filter out a malicious collective of raters which makes up almost half of the rating population. Finally, we show that the model is able to reveal provider behavior that is biased in favor of specific groups of raters.

1 Motivation for a Reputation Reporting Service

Clients choose a service provider on the basis of (1) price versus promised performance aspects such as, for example, a provider's contractual terms in an SLAs and (2) the client's confidence in how well this provider will deliver on the negotiated performance. Aggregating the confidence of many clients resembles the reputation of a provider. The importance of reputation in the clients' decisions often is underestimated in the designs of electronic market places. This has been the case in, for example, Giovanetti and Ristuccia's [6] analysis on the band-x backbone bandwidth market, where the researchers found that clients did not rely much on the reported performance numbers, but more so on the reputations of large, well-known providers.

Introducing widely accepted reputation systems for e-services requires addressing a collection of pitfalls that are inherent to them. Most of these pitfalls are not of a technical nature. For technical policies such as "we want to keep commentators anonymous", we are able to devise technical solutions. The questions that are difficult to address are in the nature of the design philosophy.

For once, one needs a definition of the scales by which to measure the reputation of a provider - a definition which includes the understanding of the pragmatic meaning of these scales. The reputation model we are presenting makes only one assumption about the choice of the reputation valuation function, namely that the reputation system operator has chosen an appropriate function.

The question, then, arises: Can we trust the rating submissions from all raters? Should we give more weight to those raters who are seemingly more trustworthy than those who might be less well-informed? To address these questions, our model has the ability to redistribute the influence it assigns to a rater on the aggregated scores, and to do so in favor of certain better informed raters. In addition, we are able to produce different views of the reputation results, depending on which rater, or set of raters, we ex ante trust more.

A separate threat model category exists on the side of the providers: Do they treat all their clients equally? Can we recognize provider discrimination? Can we distinguish discrimination behavior from biased client behavior? We address these different threat models by simulation them in challenging rating scenarios.

Finally, do clients have some incentive to submit ratings at all? Do they have a tangible incentive state their experiences

truthfully? We create these incentives by listing the clients as contributors to the reputation system and adapt policies that discourage undesired behavior.

A more detailed description of the reputation system's model, as well as further examples and demonstrations of its applicability can be found in [8]. This includes a detailed investigation of the technical model properties, as well as a practical example. In the following section we discuss some of the design choices and prior observations we made that led to the development of the reputation system as we will present it in section 3.

2 Reputation System Philosophy

2.1 Asymmetric Agents

Our system distinguishes between raters and rated entities, and allows only for uni-directional ratings. Our design requires this distinction because our market model assumes a regular business-style scenario with separate clients and providers. In this market model, clients are usually *not* providers for the same service, as in peer-to-peer networks or eBay-style trading transactions. This distinction has significant drawbacks, because it is easier to assess the credibility of raters when judging their rating behavior alongside their providing performance. For example, in peer-to-peer systems, I can distrust both the ratings submitted by a peer who has a poor provision performance, and also all the peers who rated this peer highly. In a client-server market model we lack such a high degree of interaction and rating connectivity. Nevertheless, we want to use such a model to calibrate raters' comments by the quality of comments they are submitting.

2.2 Quantitative Scales

We require that all users of our reputation system agree on one scale and assigned meaning of the reputation variable, which together characterize the recommended ordering of preferred providers. This means that the reputation variable has to be an objectively measurable, numeric value $\in \mathbb{R}$ which can be observed by all users in the same way. This $\mathbb{R}$-reputation variable may be a technically measurable metric, such as the fact that the service has a response latency of a few seconds. But the metric could also be a percentage variable which notes providers' performances in the form of a qualitative statement such as "90% of all transactions with this service were successful". With transformation functions like this, one can adapt many qualitative observations, including binary ones, to a numeric $\mathbb{R}$-variable.

Why should we not operate more with qualitative observations though? A qualitative assessment usually produces a more accurate picture, but leads to difficulties at the point of comparison and aggregation of observations. The motivation, for this requirement is that most of the times, when evaluating and judging an object - which is what reputations are about - one does so in order to make a choice, or better to produce a ranking of preferences. Thus, our desired outcome is not qualitative, but actually a quantitative description of precise orderings. In order to obtain the quantitative results, we need at some point a mapping from qualitative to quantitative observations. It is our assertion that we achieve more transparent aggregation results, if this transition from qualitative to quantitative observations is done rather earlier than later in the aggregation process.

The technical reason for demanding objectively measurable reputation input parameters is that our rating model is not able to adjust for generally biased raters, i.e. raters who judge entities in the same way as other raters, but who nonetheless assign all entities one full grade lower than the unbiased raters. The model would recognize behavior like this as the grounds for a "poor rating", and would reduce the influence that such a rater would have on the final scores. On a more general level, though, as pointed out by Pennock [10], a reputation system would need to be independent of the scales. Economists in general believe that the absolute magnitude of one user's scale cannot be compared with another user's (see Arrow [2] and Sen [11]). This is due to the fact that users inescapably are biased, and implies that users' utilities cannot be added up because they are invariant under linear transformations. Thus, it is necessary to require agreement among all the users on the scale of the reputation values. This agreement allows clients to have a certain variance with regard to the technical applications of the metric.

2.3 Rating Commentators

The reputation system calculates the performance scores of the services by averaging the comments submitted by the rating-clients. After collecting the divergences between all the comments given by a rater and the overall scored obtained, we derive a quality rating about the rater. Then the rater's quality variable is fed back into the rating of the performance scores, as a weight on his comments.

2.4 Truth Revelation Incentives

Given the uni-directional rating design of the system, we need explicit incentives which will entice clients to submit comments about their experiences with the services, and

moreover, to report on these truthfully. We create this incentive for each client by publishing, as a part of the reputation system, a list of clients who contributed comments and the number of comments they did present. We expect that a client who is listed as a rater can expect to be treated well by the service providers, because his listing is a signal to them that in future he also will submit reports on their performance. The number of comments that a client has submitted reveals whether a client, relative to his activity level, has supplied more than just a token commentary.

To entice truth revelation from the clients, we have to apply another measure to our reputation system. We considered the possibility of publishing the rater reputation that our system calculates in the process of aggregating scores. There, a rater's reputation reflects his rating accuracy and the influence assigned to his comments by the reputation system. Making this rater reputation public would increase the rater's incentive to provide high ranking commentary, because this would send a stronger signal to the providers. However, the rater reputation should be kept undisclosed by the reputation system, since knowing this would also create an incentive for the clients to modify or even fabricate their comments in a targeted way. They could do so by submitting comments that reflect the currently published scores for the services. So actually, in order to promote truthful comment submissions, the reputation system adopts a policy of striking clients off its list of raters if they submit erratic or modified comments. We can recognize such clients from their particularly low rating reputations. We should point out that the clients' reputations and their influence on the ratings will decline and remain unnoticed outside of the reputation system before we strike them off the rater register. We strike them off once the coherency of their ratings underpasses a certain low standard. Figure 3 in section 4.1 gives an example for how inaccurate raters can be identified through the reputation system. With this policy, raters have no incentive to submit untruthful comments. Instead, they have an incentive against this, particularly because they are not able to gauge how high their own rating reputation is. Rating reputations depend on all the other comments and thus are only known by the reputation system.

Another advantage in adopting this policy of striking off random or malicious raters is that doing so reassures providers that the reputation system is meaningful and that they should accept the published rankings. This acceptance should increase their willingness to improve their service quality where necessary. Service quality improvements should actually be self-evident if the reputation system creates competition over quality aspects.

3 The Rating Mechanism

We have a set of l *raters* $R = \{r_1, .., r_l\}$ who rate k *service providers* $P = \{p_1, .., p_k\}$ on the performance they experienced as clients. The providers perform at a certain 'real' performance level and this is the value we ideally would like to obtain as a final score for this provider in the reputation reporting system. However, this 'real' level is an unknown variable, possibly even unknown to the provider himself. Every time a client requests services from this provider the client experiences a probabilistic value of this performance. However, these individual 'experiences' may be observed slightly differently by each individual client. The individual differences arise from a number of reasons such as them applying different measurements, or aggregating their experiences on different scales.

Every so often, at time t rater $r_A \in R$ will make a *comment*: $c_{t,\, r_A \mapsto p_B} \in C$, about provider $p_B \in P$, reporting his observations in form of a numerical value: $val(c) \in \mathbb{R}$. The following calculations are performed for all raters and providers, however in order to simplify the notation, we will demonstrate the calculations for an exemplary rater r_A and an exemplary provider p_B. As the comments are collected at a central point in the reputation system, we define all comment time stamps t to be distinct. Raters can submit more than one comment about a provider and these can be distinguished by the time the comment was made.

After collecting a sufficient number of comments[1] on a provider: $|\{c_{t,\, r_* \mapsto p_B}\}|$, we can calculate an initial score for this provider. In our terminology '$*$' represents any client or provider applicable. Before entering the iterative loop, we initialize the provider scores to obtain the *initialization vector* $s^{(0)}_{p_{1,..,k}}$. This initialization simply averages the values of all the comments:

$$\forall\, p_B \in \{p_1, .., p_k\} : \quad s^{(0)}_{p_B} = \begin{cases} \dfrac{\sum\limits_{c \in C'} val(c)}{|C'|}, & C' = \{c_{t,\, r_* \mapsto p_B} \in C\} \\ 0, & \textit{if} \quad C' = \emptyset. \end{cases} \tag{1}$$

C' is the set of all comments made by any rater about provider p_B and if this set is empty, we assign a score of null. The (0) represents the iteration number[2], indicating the initialization procedure at this point, where we simply average the received comments.

[1] It is possible to develop a confidence value that indicates how many comments are sufficient to achieve statistical integrity. However, we decided to omit developing and analyzing such a variable.

[2] In our notation, where variables (e.g. s) are indexed by the iteration number (e.g. n), such as $s^{(n)}$, the iteration number is kept in parenthesis in the superscript in order to distinguish these from power operations.

Next, we calculate the score each rater would assign by himself for each provider. For each provider, we take the collection of raters having submitted comments on this provider and then for each rater within this set we average his comments on this provider to obtain a local score $local.s_{r_A \mapsto p_B}$. Here, C' is the set of all comments made by rater r_A at different times about provider p_B and if this set is empty, $C' = \emptyset$, we assign a score of null, $local.s_{r_A \mapsto p_B} = 0$:

$$\forall\, r_A \in \{r_1, .., r_l\} : \forall\, p_B \in \{p_1, .., p_k\} :$$
$$local.s_{r_A \mapsto p_B} = \begin{cases} \dfrac{\sum\limits_{c \in C'} val(c)}{|C'|}, & C' = \{c_{t,\, r_A \mapsto p_B} \in C\} \\ 0, & \textit{if} \quad C' = \emptyset. \end{cases} \tag{2}$$

After calculating initial and local scores for all providers, we enter the iterative loop and calculate for all the raters their reputations $q_{r_A \in R}$: The reputation is calculated from the difference between rater A's comments and the above calculated initial scores $s_p^{(0)}$. Then we calculate the standard deviation of this difference. Of this difference we take the norm by dividing it by the number of comments submitted by this client. This yields our reputation value, except that, as it stands, a good rater with a high rating reputation would be assigned a smaller value q. In order to correlate good reputations to high reputation values, we invert this final reputation value. C' is the set of all comments made by rater r_A about any provider and if this rater did not submit any comments at all, he is assigned a rater reputation of null, $q_{r_A} = 0$:

$$\forall\, r_A \in \{r_1, .., r_l\} :$$
$$q_{r_A}^{(n)} = \begin{cases} \dfrac{|C'|}{\sqrt{\sum\limits_{c \in C'} \left(s_{p_*}^{(n)} - val(c)\right)^2}}, & C' = \{c_{t,\, r_A \mapsto p_*} \in C\} \\ 0, & \textit{if} \quad C' = \emptyset. \end{cases} \tag{3}$$

Following, we use the raters' reputation values as a weight on their comments, and therefore obtain an updated score for the provider ratings.

This weight each rater obtains is directly proportional to their reputation value q, and is represented as their share of influence $infl$ on the reputation scores. From the above collection of all raters $\{r_1, .., r_l\}$ of the providers $\{p_1, .., p_K\}$, we take their reputation values $\{q_{r_1}, .., q_{r_l}\}$, and divide 100% of available influence into slices $\{\textit{infl}_{r_1}, .., \textit{infl}_{r_l}\}$, proportionally to these reputation values. Additionally we introduce what we will label the rating model's *selectivity weight* variable w, which reinforces the selective effect of the model for $w > 1$ and dampens the model effect for $0 < w < 1$. For $w = 0$ the model effect is entirely disabled and the resulting scores are result of plainly averaging the input comments. This selective weight variable is essential to the utility of the rating model, as we need it to adjust the model's balance between selection and inclusiveness to any certain market scenario. In case the model was run on an empty input set without any comments submitted to the system, we assign all influence values to null, $\textit{infl}_{r_A} = 0$:

$$\forall\, r_A \in \{r_1, .., r_l\} :$$
$$\textit{infl}_{r_A}^{(n)} = \begin{cases} \dfrac{\left(q_{r_A}^{(n)}\right)^w}{\sum\limits_{i=1}^{l} q_i^w}, & C \neq \emptyset \\ 0, & C = \emptyset. \end{cases} \tag{4}$$

Finally, we compute an updated performance score rating $s_{p_B}^{(n+1)}$ for the next iteration $n+1$ by summing up the averaged comments from 2 multiplied with the influence shares from the previous equation (4). Note that if in the previous equation (4) all influence values have been set to null, $\textit{infl}_{r_A} = 0$, the resulting scores automatically also are all null:

$$\forall\, p_B \in \{p_1, .., p_k\} :$$
$$s_{p_B}^{(n+1)} = \sum_{i=1}^{l} \left(local.s_{r_i \mapsto p_B} \times \textit{infl}_{r_i}^{(n)} \right). \tag{5}$$

This concludes iteration n, the next will continue at step (3), with these now updated provider scores and raters' influence values. This iterative process will continue, until the updated differences of the calculated performance scores fall below a convergence threshold δ. Convergence is usually very rapid, and we found a fixed threshold of $\delta = 0.0001$ to be practically suitable in most conceivable scenarios. The algorithm is robust such that if one runs this rating algorithm with no or only one comment as input, the algorithm converges immediately in the first iteration.

3.1 Confidence Value

If only a subset of raters has submitted ratings on a particular provider, the resulting score of the provider is only based on the comments from these raters. This may mean that a resulting score can be derived from raters who have a low overall rating reputation. There is little to be done about the score in such a case, since there are no other ratings that we have more confidence in. However, from equation (5) we can calculate the sum influence values that we used to compile a resulting score and therefore obtain a confidence value on this score. The confidence value $\textit{confidence}_{p_B}$ is calculated for every provider, after the algorithm has converged:

$$\forall\, p_B \in \{p_1, .., p_k\} :$$
$$\textit{confidence}_{p_B} = \sum_{j \in R'} \textit{infl}_j,$$
$$R' = \{r_* \in R \,|\, \exists\, c_{t,\, r_* \mapsto p_B}\}. \quad (6)$$

The confidence value $\textit{confidence}_{p_B} \in [0, .., 1]$ is a measure that states how many per cent of the total accumulated rater reputation participated in deriving the score for this provider p_B.

3.2 Multiple Convergence Points

The algorithm converges very well, but is able to produce more than one convergence point. In this section we first present the convergence properties of the algorithm and then explain the implications of multiple convergence points.

For a scenario of non-convergence to exist at all, it is necessary that in the course of the algorithm's iterations, we reach in equation (5) an influence vector which is identical to a vector reached in a previous iteration. Otherwise, the shifting of influence shares will progressively and eventually converge to a stable set of scores, because the local scores $local.s$ are fixed after the initialization and the algorithm cannot avoid convergence without traversing the same influence share vector twice. It is hard to construct a scenario that leads into a an infinite cycle, the most promising approach would be to devise a scenario that fulfils the arrow impossibility theorem [2]. However we did not encounter any such scenarios and in general the algorithm converges rapidly, within several dozen iterations. Only when the solution approaches discontinuities, when the solution falls onto an altogether different convergence point, the convergence can take up to hundreds of iterations. The convergence properties are investigated in greater detail in [8].

Depending on the setting of the selectivity weight factor w, any given scenario set of comments has one unique or multiple possible sets of scores to converge to. In fact, for $w = 0$ all scenarios have exactly one convergence point, this being the simple average of all comment values, and for a very high setting of w, as many convergence points exist as there are raters. To reach all these convergence points it is necessary to replace the initialization vector $s^{(0)}_{p_{1,..,k}}$ of equation (1) by the values of the comments of each rater. However, as soon as a second convergence point exists, the results from the reputation model are questionable. The intuition behind this is, that if for the same comments and selectivity weight, we are able to obtain more than one possible outcome, there is no model-inherent way to determine, which of the outcomes to choose. The motivation for choosing simple averaging as an initialization vector in equation (1) was to speed up the convergence. Therefore we suggest to limit the settings of w for to the range of the unique solution scores, unless one has an application suitable method for selecting one of the convergence points over the others. One such method could be to user customize the scores and to choose the convergence point from the perspective of this user's comments.

If starting from the same averaging initialization vector, and continuously changing the the setting of the selectivity weight factor w, usually also results in continuous shifts of the convergence point. Commonly, when changing from one to the next convergence point, transitions are continuous. However, discontinuous jumps are possible if the two convergence paths are connected by discontinuous gradients. In our extreme example the two voting blocks correspondingly have two convergence points, and scores gravitate to either of the two voting blocks, as long as w is set strong enough to promote such selectivity. In this example, between $w = 1.77$ and $w = 1.78$ the convergence assumes a slightly different slope and then changes the direction of the convergence slope. Figure 1 and 2 show the distribution of influence shares changing from iteration to iteration for the two borderline values of $w = 1.77$ and $w = 1.78$ and display how this transition comes to place. The very same clients who dominate in setting A, are marginalized in setting B and vice versa.

4 Threat Model Analysis

The scenarios in this section contain 100 raters and 20 providers. In these scenarios, the providers are set to perform at a certain average performance level, which is characterized by a numeric value. The goal of our rating model is to identify these performance levels from the comments submitted by the raters. The scenario assumes that raters are not able to capture the performance level of a provider perfectly, because performance is compiled from the quality of interactions over a time period. But these scenarios assume that the providers will on average perform to a certain measurable level, therefore comments are drawn from a random variable with a normal distribution where the mean equals this performance level. While each provider will have a different performance level to show the figures more clearly, one should note that the choice of the actual level is irrelevant to the the rating model, only the deviations from this level are relevant.

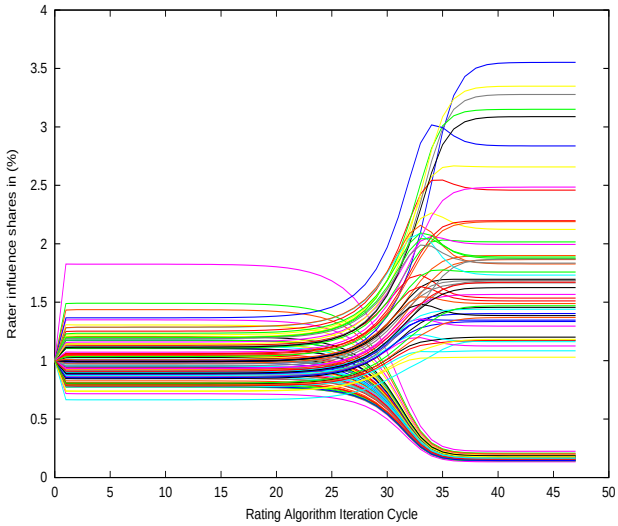

Figure 1. Slope of influence shares for the divergent example at $w = 1.77$, converging to voting block A.

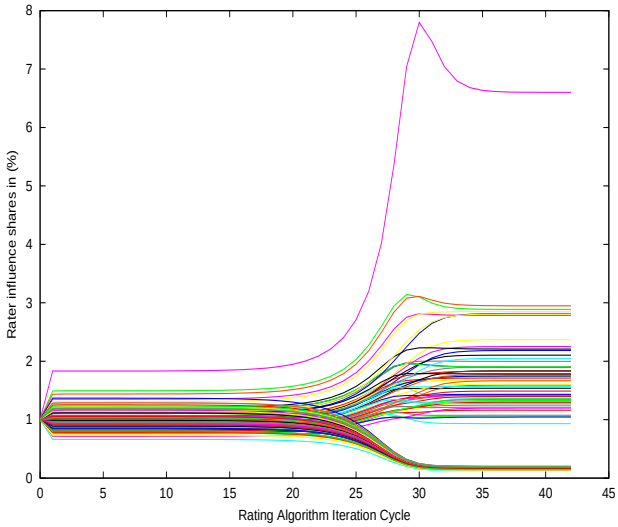

Figure 2. Same example during convergence with $w = 1.78$ and gravitating to voting block B.

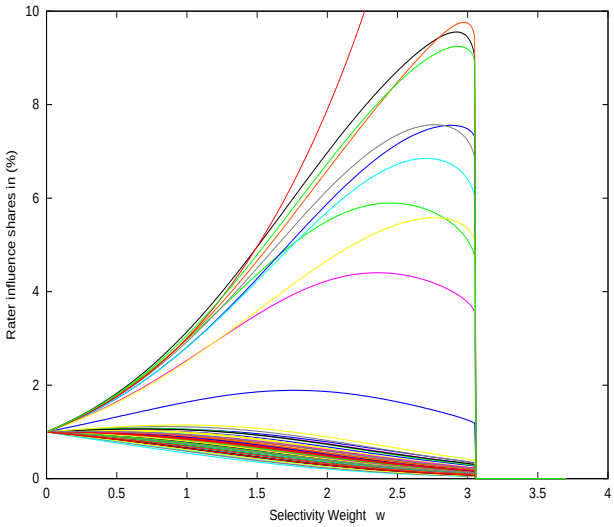

Figure 3. Influence Shares for a Mix of Ten Raters with $\sigma = 0.25$ and 90 Raters at $\sigma = 1.0$.

4.1 Different Rating Accuracy

We want to confirm if the rating model is able to discern raters with a weak rating consistency and set up a scenario with sets of raters of different rating accuracy. One group of 10 raters has a low standard deviation of $\sigma = 0.25$ and a second group of 90 raters has a relatively high standard deviation of $\sigma = 1.0$. In figure 3 we can identify the two distinct bunches of graphs, the top bunch corresponding to the group of low deviation raters and the lower bunch relating to the ten raters with the high deviation. In this scenario, one would choose a selectivity factor of about $w = [2.5, .., 3.0]$ in order to put the balance in favor of the ten raters with a low deviation. In doing so, we would ensure that the scores are only drawn from raters with the higher accuracy.

4.2 Malicious Rater Collective Manipulating One Provider's Score

The most important threat model of "bad" ratings is a malicious collective of raters who attempt to influence the score outcome in a coordinated and directed fashion. For such a scenario 40 raters apply a deviating commenting agenda by adding an offset of -2.5[3] to their comment values. If this set of deviating raters would apply their rating bias to all

[3]The offset value of -2.5 was chosen such that one could visibly recognize the impact of the deviators on the scores, with the combined scores effectively lowered by an offset of -1. One could also question if a relatively smaller offset value for such a deviation would amount to a threat model or would have to consider such comments valid and rightfully include these in the scores.

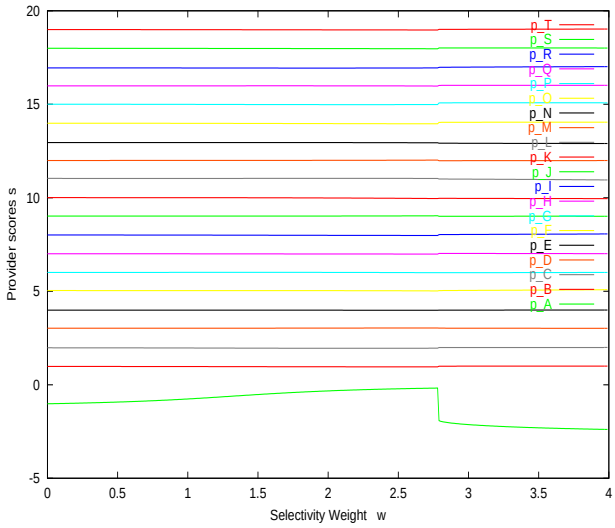

Figure 4. Scores, where all raters agree on all ratings, except for 40%, who disagree on one provider, p_A, and apply a rating offset of -2.5. For $2 < w < 2.7$ this offset is canceled out by the reputation system.

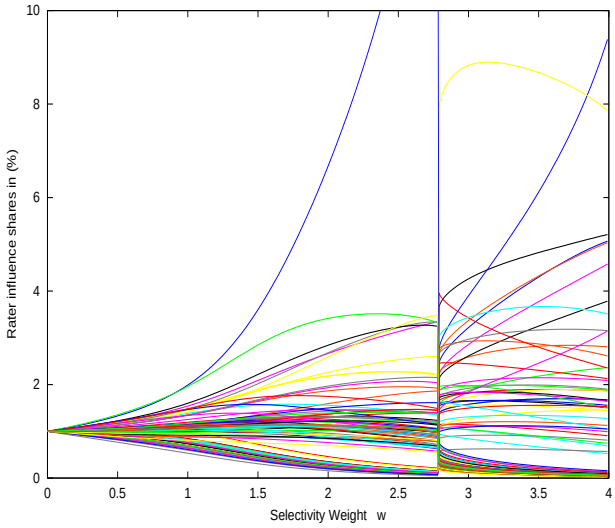

Figure 5. Influence shares for the scenario of figure 4. The 40% raters who apply the rating offset are given by the reputation system a distinctly lower influence in the range of $2 < w < 2.7$ than the other raters.

the rated providers, it would be easier to filter them out. As it is much harder to identify and deselect these deviators if they attempt to artificially alter the rating of only one of the providers, the bias is applied only to provider p_A. In fig. 4 we see how the scores are effected by this deviation, because for $w \leqq w < 0.5$ the scores are lowered by one unit from the "correct" values. Increasing w makes the model more selective and eventually for $w > 2.5$ the rating scores reflect the rating pattern of the 60 "good" raters only. From figure 5 we see though, that the gap between the graphs of the 40 deviating raters (represented by the graphs close to 0 in the range of $2 < w < 2.78$) and the 60 other ones is very narrow. This shows that the rating model is pushed to its limit, although it still does achieve the goal of removing the effect of the deviators for a selectivity weight value of about $w > 2.5$.

At this point we would like to determine an appropriate setting for w in this scenario, such that deviators are rightfully disabled, but the other raters form a meaningful aggregation. Therefore we search for a second convergence point by replacing the initialization vector in equation (1) with the comment values of the deviators. Figures 4 and 5 actually show the model results drawn from the altered initialization vector that is favoring the deviators. At $w = 2.78$, the second convergence point is available, the scores follow a discontinuous jump to the scores of the deviator group and the influence distribution is reversed for both groups. With the existence of the second convergence point, the solution is not unique and it could be debatable which of the convergence points should be chosen. Therefore we choose to disregard results that offer two convergence points.[4] Within the range of possible unique model solutions, we want to maximize the rating model's selectivity effect in order to enable the model to filter out the deviating raters. Choosing $w = 2.78$ satisfies both of these criteria and yields us model solution scores that are practically free of influence taken by the deviators, which at this point collectively only amounts to 4.41%.

4.3 Deliberately Inconsistent Provider Performance

The scenarios we discussed in the previous sections assume that providers display a consistent performance behavior. This does not imply that they have to deliver identical performance every time, but we assume that clients base their ratings on a number of transactions with a service and that

[4] We can choose to take into account results that allow for more than one convergence point, if we have a method for choosing the desired initialization vector to reach the desired convergence point. One such method is to take a client-specific view, and select the convergence point according to this client's comments.

statistically the clients experience on average a similar performance from a provider and that this average is characteristic for this provider. Deriving a reputation with the model under such assumptions for a provider who displays a high variability in his performance would still result in the scores correctly reflecting his average performance.

However, what happens if a provider does not deliver a varied performance by statistical chance, but rather has opted to treat a specific set of clients with a specifically worse or better performance level? For example, provider p_A could decide that he delivers a guaranteed-level 99.99% of the agreed performance terms to a majority of 75% of preferred customers and delivers as low as only 50% of the agreed performance terms to the minority of 25% of the remaining contracting clients, when he needs to use these clients as a best-effort-serviced buffer for his performance demand variances. With the clients then submitting their observations as comments, this would result theoretically in a performance score of 87.49%, if applying a plain average. We would argue that this does not represent the performance of the provider well at all. Such a result neglects the good and the bad of this provider's strategy.

In order to make this scenario statistically more realistic and relevant, the majority group experiences a normal distribution with no performance above 100% and a very low standard deviation of $\sigma = 0.25$, simulating the guaranteed performance level. The minority group's normal distribution of comments on the other hand will never fall below 50% and displays an extremely high standard deviation of $\sigma = 25.0$, reflecting the best effort characteristic of their service experience.

Figure 6 compares the model's score outcomes for the two possible convergence points of p_A's rating and practically, with the statistical circumstances mitigating p_A's performance bias, a plain averaging of comments results in a flattering rating score of 91.4%. The remainder of this scenario contains again 100 raters, with 25 reporting the lower performance and a total of 20 providers, where the other 19 all behave consistently and predictable. While the first convergence point is found from the overall average as the initialization vector, the second is derived from an initialization vector formed from the comments the minority group is submitting. As in the previous scenarios, as soon as the selectivity weight is high enough such that the resulting scores are not compromising between the two groups anymore, the second convergence point appears, that is for $w \geqq 1.85$. At that point we can clearly identify the two general performance levels (99.99%; 50%) this provider is supplying to the two groups of clients. We can individually identify which of the clients belong to either of these two groups by observing the exact reversal of the influence shares held by either group for the different convergence points. Ironically

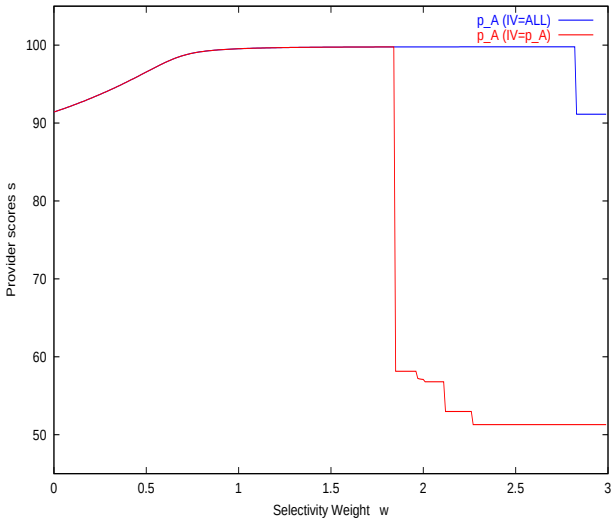

Figure 6. Scores for both convergence points where provider p_A is deliberately delivering inconsistent performance.

though, when increasing the selectivity weight, eventually for $w \geqq 2.83$ the resulting scores center on one of the minority group raters who happens to be very close again to the plain average which we obtained with $w = 0$.

From the point of view of the scores, this scenario is identical to a set of clients with a rating bias, as we described in section 4.2. By analyzing the submitted comments it is not possible to distinguish between a provider actually treating a client differently than others or a client turning out a different rating than the performance he has received. The model however is able to reveal these divergences, particularly if these are part of a general pattern and not random flukes. However, to tell which of the both cases of deviance we are dealing with is external to this rating model and needs to be resolved with other methods.

5 Related Work

The idea to introduce reputations for raters in our reputation aggregation algorithm was inspired by the way page rankings are calculated in Google (transfer of endorsement[9]), and has been picked up by other reputation system research in various ways. Chen and Singh[4] integrated this in their versatile reputation system, which allows for either bi-directional or unidirectional ratings and is able to integrate plain text comments from raters. Chen and Singh's research is closely related and supersedes ours in its functionality. However, our algorithm and the threat models are more specifically tailored to the uni-directional recommen-

dation case of client-server oriented business scenarios.

Collaborative filtering systems are reputation systems that use similarity-based approaches to provide a function that aggregates recommendations of other clients in relation to a specific client's preferences. Pennock et. al. [10] analyze the theoretical foundations of collaborative filtering systems under the aspect of social choice theory. Dellarocas [5] presents a cluster filtering algorithm, which he evaluates in terms of its ability to eradicate the effects of unfair ratings.

For peer-to-peer reputation systems, Kamvar et. al. [7] develop the EigenTrust algorithm, which uses a similar trust propagation concept, transfer of endorsement, which recursively feeds the trust information back into the aggregated scores. Given that EigenTrust has been designed to suit a more specific application than our reputation system, it is interesting to note that both are able to identify similar sized proportions of malicious peers.

Bayesian estimation is another approach to classify sets of agents by their behaviour [1][3]. Buchegger [3] develops a routing protocol for mobile ad-hoc networks, which uses a modified Bayesian estimation procedure to eliminate false information among the second-hand reputation information and to isolates misbehaving nodes. The main difference between our approach and Buchegger's is that Buchegger calculates the reputation aggregation from the viewpoint of an individual node, where our approach is the one of a service that does not use first-hand information. Being able to use first-hand information strengthens the filtering of misinformation, but it always makes the resulting recommendations partial to the collector of the first-hand information.

6 Financing the Operations of the Reputation System

Operating such a reputation system needs to be funded in some way. However, the way we choose to fund the reputation system has implications on the incentives put forth by the reputation system. With an reputation system we aim to resolve some of the market inefficiencies, which presumably would yield benefits that could be tapped for refinance. Clients benefit directly from the higher reliability of services from better providers, and the providers benefit from the market's ability to drive "lemon"-providers out of business and the therefore generally higher confidence by the clients in the services market. Some providers may benefit by charging a premium for delivering service with higher reliability.

Possibly the easiest route to obtain the funds to run the reputation service would be to demand subscription fees from the listed providers, as these could afford to write off the fees as a form of advertisement for themselves. The other incentive for the providers to pay for a subscription would be to lock foul competitors out of the market, as they now would be identifiable through the reputation system. Effectively they would be forming a guild of "accredited" providers and thereby maintain a clean market for the clients. However, we would choose against funding the system by the providers, as this would create the wrong incentives for the operation of the reputation service. The operators of the reputation system would lack the incentive to promote transparent performance comparisons between providers and thereby promote competition among them, but would have an incentive to set a high bar for the level for market entry. Our reputation system seeks to aid the clients' choice to make the market in general more efficient, and a provider funded system would naturally follow different goals.

The economically cleaner solution would be to demand payment from the clients and thereby provide a direct incentive for the reputation system to cater to the needs of the clients' choice. A client funded reputation system does not have an incentive to alter the objectivity of its provider reputation rankings and if it is subject to competition itself, where clients choose a reputable reputation system, the reputation system also has a disincentive to accepting bribes (or advertisement) from the providers. However, demanding clients' to pay is likely to meet their resistance, as it can be expected that some of the clients rather not use such an advisory service at all than having to pay for it, even if such could be shown to pay off in the long run. Moreover, clients could undermine this revenue model by forwarding or even publicizing the reputation system's ranking lists. Further, the actual pricing model would be a sensitive issue, as a fixed fee might be too high for some small one-time client and yet negligible for a high-turnover one. Addressing this through discriminatory pricing schemes is likely to introduce distortions and using micro-payments that scale proportionally with the volume of use is likely to introduce unreasonably high transaction overheads. A specific problem is system startup, as the value of the reputation system increases along with the number of gathered clients' review comments. Therefore, one could not use the client funding to cover the startup phase. On a general level, it may turn out that the clients' individual per-transaction-profit from using the reputation system is not sufficient to cover the running cost of the system, although amortized they themselves or the whole market would profit. This is not an unlikely case and one common solution is to call for a government sponsored funding solution.

Where a system's net benefits need to be amortized either over time or aggregated over a number of market participants, it is common to seek for government sponsoring, as

the system meets the criteria to be considered an infrastructure element. It is our anticipation that the reputation system would be such a candidate due to all the complications stated in our discussions about client funding and rejecting provider funding on the ground of incentive incompatibility.

Another possible solution is the mix of government and client funding. In that case one would draw on client charges for enquiries about high stakes services that yield client benefits clearly outweighing the charges and government funds for the remainder, in particular any startup periods. In the UK's energy market, the government supports such price comparisons in order to make the market more dynamic and transparent. Even in the absence of direct government funding for a reputation system, it may well be necessary to introduce government regulations that protect the reputation reporting system. Otherwise providers not only can lack cooperation with such a reputation system setup, but could actively obstruct its work, for example by demanding from their clients not to disclose their experiences that stem from using the provider's service. Such obstructions would most likely be expected from providers who have built their reputations through other means, such as advertising, or sheer size and now attempt to protect their advantageous profile.

7 Conclusions

At a pragmatic level, we showed that our reputation aggregation system is able to counter three hard threat scenarios. The algorithm highlights the difference between raters with a high and a low rating variance and is able to shift influence to the raters with a low rating variance, even if the vast majority of raters has a high variance. Further, we show that the model is able to filter out a malicious collective of raters, even if this group singles out only one provider and is close to being the majority. The third model demonstrates how the model can be used to identify unfair behavior on the part of the providers.

We were able to demonstrate that the rating model is able find the largest subgroup of raters with the closest agreement and through this to filter out various forms of undesirable rating behavior. Most commonly, we recommend to set the selectivity weight variable w to the maximum setting that leads to a unique solution (ones that have only one convergence point). However, it is possible to adapt the reputation model's aggressiveness as a selective device to application-specific demands, which is explained in greater detail in [8].

The reputation aggregation model is very versatile, and so has been applied to the academic peer review process. Chapter 8 in [8] demonstrates on the example of a conference review, how the model is able to numerically capture the notion that some reviewers are better at contributing to consensual review results than others. We show how the review results can be gracefully adjusted to reflect the quality of the reviewers, without lowering too much the confidence value of the resulting review scores. This example also displays the scalability of our reputation model, which is able to produce useful scores even when applied to a small numbers of entities with peculiar characteristics.

References

[1] K. Aberer and Z. Despotovic. Managing Trust in a Peer-2-Peer Information System. In *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM)*, 2001.

[2] K. J. Arrow. *Social choice and Individual Values*. Yale University Press, 1963.

[3] S. Buchegger. Coping with Misbehavior in Mobile Ad-hoc Networks. PhD Dissertation, Ecole Polytechnique Fegerale de Lausanne, January 2004.

[4] M. Chen and J. P. Singh. Computing and using reputations for internet ratings. In *Proceedings of the 3rd ACM conference on Electronic Commerce*, October 2001.

[5] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the ACM Conference on Electronic Commerce*, pages 150–157, Minneapolis, Minnesota, USA, 2000.

[6] E. Giovannetti and C. A. Ristuccia. Estimating Market Power in the Internet Backbone Using Band-X data. Cambridge Working Papers in Economics, Department of Applied Economics, University of Cambridge, June 2003.

[7] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the Twelfth International World Wide Web Conference (WWW)*, Budapest, Hungary, 2003.

[8] J. H. Lepler. Cooperation and Deviation in Market-Based Resource Allocation. PhD Dissertation - http://www.cl.cam.ac.uk/TechReports/UCAM-CL-TR-622.html, Cambridge University, November 2004.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[10] D. M. Pennock, E. Horvitz, and C. L. Giles. Social choice theory and recommender systems: Analysis of the axiomatic foundations of collaborative filtering. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, 2000.

[11] A. Sen. *Handbook of Mathematical Economics*, volume 3, chapter Social Choice Theory. Elsevier Science Publishers, 1986.