Credibility-Based Trust in Social Networks

Noel Sardana[1], Robin Cohen[1], Jie Zhang[2], and John Champaign[1]

[1] Cheriton School, University of Waterloo, Waterloo, ON, Canada N2L 3G1
[2] Nanyang Technological University, 50 Nanyang Ave., Singapore 639798

Abstract. In this paper, we develop an approach for trust modeling that uses expectations on probability distributions to derive the predicted benefit of a message in a participatory media setting. Our solution addresses the issue of false information propagation in social networking settings by examining the role that global credibility might play when determining peer trustworthiness. We present an overall algorithm, illustrated with examples and discuss next steps. The model is presented in the context of related work on trust modeling and on selecting messages to present to users in peer-based networks.

1 Introduction

Today, users are increasingly engaged with online media via the internet. A plethora of available information, ranging from online news networks to various forums to social and participatory media sites, contributes to information overload germane to our information rich society. Moreover, the advent of Massive Open Online Courses (MOOCs) and the increasing use of both online retailers (like eBay) and information feeds (like Twitter) suggests the following question: how can we help users sift through excess information to retrieve the most relevant objects of interest? In this work, we offer a novel approach that copes with false information propagation by incorporating an element of global credibility.

Related work in trust modeling that motivates the development of our approach includes: Jøsang et al.'s Beta Reputation System [5], Teacy et al.'s TRAVOS system [11], and Zhang et al.'s Personalized Trust Model PTM [12], as well as a Bayesian Credibility Model (BCM) by Seth et al. [10] and a model for recommending annotations of learning objects to peers, of use in large social networks, termed Learning Object Annotation Recommendation (LOAR) [1]. We present a brief overview of these models, below.

Beta Reputation System (BRS): The Beta Reputation System (BRS) [5] for use in e-marketplace reputation systems, provides a foundation for later trust modelling work carried out by Zhang et al. [12]. BRS is foundational because it is grounded in probability theory; it uses the beta probability distribution, which is a conjugate prior for binary events. In particular, in BRS, the expected value of a beta density function is given by the formula

$$E(f(p)) = \frac{\alpha}{\alpha + \beta} \tag{1}$$

In [5], the authors interpret this expected value as the probability of some positive outcome occurring in the future, where $\alpha = r + 1$ (r being the number of positive outcomes that occurred in the past) and $\beta = s + 1$ (s being the number of negative outcomes that previously occurred). Thus, the expected value of a beta distribution is a suitable trust metric. One characteristic of BRS, however, is that all ratings, even from potentially dissimilar peers, are treated equally.

TRAVOS: Teacy et al. developed a system called TRAVOS to model trust relationships between agents in virtual organizations [11]. Like BRS, TRAVOS defines a trust metric to be the probability that a trustee will perform on a future obligation and uses a beta pdf to model relative trust probabilities. Unlike BRS, TRAVOS incorporates the notion of confidence in the inferred trust metric, i.e., it models the probability that the true trust metric lies within a certain margin of error. TRAVOS also allows the truster to seek third party advice when this probability is below some threshold. Trusters can then aggregate reports to derive a more confident perspective. TRAVOS, however, relies on the assumption that truster and pundits have extensive historical dealings that enable each pundit's expected honesty to be assessed. There is also no time discounting of reports.

Personalized Trust Model (PTM): Zhang and Cohen [12] propose a personalized trust model to determine whom to listen to amongst a network of buyers and sellers in an e-marketplace domain. Global advice from other buyers (advisors) is combined with the buyer's own local experiences with a seller. The PTM global metric is further broken down to combine public and private trust estimates of advisors. As with other trust models, the ratings of a seller are binary. The beta pdf is then used to estimate the probability that an advisor will provide a fair rating to a buyer. To estimate the private reputation of an advisor, a, PTM defines

$$R(a)_{private} = E(Pr_a(\text{fair rating})) = \frac{\alpha}{\alpha + \beta} \tag{2}$$

In essence, the number of times the buyer and advisor have the same rating of the seller (and the number of times they are dissimilar) forms the basis of the calculation of α and β; in addition, this comparison is done within limited timewindows, to ensure timely evaluations. The advisor's public reputation is calculated similarly, measuring whether its ratings correspond to the average rating vector over all ratings of the seller. In the end, the trust level of the advisor is derived by calculating a weighted average of private and public trust (including a forgetting factor to discount less recent ratings). The trust value of the seller in turn involves a weighted combination of both private and public reputation ratings. With PTM, however, when advisors are dissimilar to a particular buyer, their advice is simply not regarded as highly (and the model does not take advantage of their advice).

Bayesian Credibility Model (BCM): Seth, Zhang, and Cohen [10] propose a Bayesian model to derive the crediblity of messages within a social network of peers for the purpose of recommending participatory media content (e.g., blog posts, consumer product reviews, Twitter tweets, etc.) to users. BCM uses the *strength of weak ties* hypothesis from social network theory to categorize clusters of users within a social network, G. The topic-induced subgraph of G, denoted G_t, is a subgraph of users who are interested in some topic, t. Users within G_t can be categorized as belonging to particular clusters, i.e., subgraphs of users that are strongly tied and affect knowledge propagation throughout the cluster in certain ways. Clusters are connected together via weak ties to form the topic-induced subgraph G_t.

For each user $u_i \in G_t$, BCM derives a topic-specific crediblity score for each message m_k, denoted $C_{k,t}$. $C_{k,t}$ depends on Contextual (how easily a message is understood, CN) and Completeness (the depth and breadth of media content, CM) information. Context and Completeness are in turn dependent on four sub-crediblity types (evidence variables):

Cluster credibility (denoted $s_{i,k,t}$) is the credibility the cluster of user u_i (denoted V_{it}) assigns to message m_k authored by some other user, u_j.

Public crediblity (denoted $p_{k,t}$) is the credibility that the entire network of users in G_t assigns to message m_k.

Experienced credibility (denoted $e_{i,k,t}$) is the credibility that u_i assigns to message k based on u_i's past experience with the author of m_k, viz., u_j.

Role-based credibility (denoted $l_{i,k,t}$) is the credibility u_i assigns to m_k given that u_j has some role (and thus has some level of expertise).

BCM has some drawbacks, including its requirement of explicit connections between users (e.g., in the form of “friendships” or some other such designation) to derive “clusters”, and the requirement of several ratings in common to derive an adequate Bayesian network.

Learning Object Annotation Recommender (LOAR): In [1], Champaign et al. draw inspiration from Zhang's PTM [12] to develop a recommender system that selects annotations on a learning object to display to students in an online learning environment. The model displays those annotations with the highest predicted learning benefits.

When viewing learning objects, students are allowed to vote on the attendant annotations (ratings are binary). The “curent” student experiences a learning object with annotations displayed in a customized fashion according to their predicted learning benefit. An annotation's predicted benefit is calculated using a combination of the annotator's reputation and explicit ratings the given annotation has recieved. An annotator's reputation is derived as follows:

1. An author q has created a set of annotations $A_q = \{a_1, \ldots, a_n\}$, each of which has an associated set of ratings $R_{a_i} = \{r_1, \ldots, r_{m_i}\}$ left by some number, m_i, of students who have previously experienced the annotation.

2. Compute a set of average ratings, $V = \{v_{a_1}, \ldots, v_{a_n}\}$, corresponding to each annotation using the associated rating set, i.e., $v_{a_i} = \frac{1}{m_i} \sum r_i$
3. The annotator reputation, T_q, is the mean average rating, i.e., $\frac{1}{n} \sum v_{a_i}$.

Here, parallels between Champaign's annotator model and Zhang's trust model begin to emerge. An annotator in LOAR corresponds to a seller in PTM, and the total annotator reputation, T_q, is akin to a seller's global reputation, $R(S)_{global}$. Moreover, student peers in LOAR act as advisors in the system, albeit the calculation of $R(S)_{global}$ from advisor experiences in PTM requires additional work to derive the trust values between the buyer and advisors, and more recent experiences are weighted higher in PTM due to a forgetting factor. Another difference is that LOAR models the predicted benefit (or trustworthiness) of *annotations*, whereas PTM models the trustworthiness of *sellers* (not products). Even if the annotator is highly regarded in a community, any one particular annotation will be influenced more heavily by the ratings it receives. The annotator reputation serves only as a proxy for ratings when an annotation has not received votes (e.g., is relatively new).

A "local" annotation reputation depends on the number of votes it receives. In particular, votes for and against an annotation are weighted according to the similarity between the current student and peer voter, which, as in PTM, is calculated according to prior votes the pair have cast in common (ignoring the complexity of time windows and focusing solely on common items). The global and local annotation reputations are then combined in one of two ways to derive the predicted benefit for the current student:

1. Using a Cauchy CDF, pred-ben$(a) = \dfrac{1}{\pi} \arctan \dfrac{vF_a - vA_a + T_q}{\gamma} + \dfrac{1}{2}$, where vF_a and vA_a are the number of votes for and against the annotation, each weighted according to voter similarity to the current student.
2. Weighting T_q as a proxy for the annotation's reputation according to some requisite minimum number of votes. That is, pred-ben$(a) = \min\left(1, \dfrac{|R_{a_i}|}{N_{min}}\right) \cdot V_{a_i} + \max\left(0, 1 - \dfrac{|R_{a_i}|}{N_{min}}\right) \cdot T_q$.

LOAR is a good first step in applying trust modelling in the online education domain. However, the heuristics of which the model makes use are not grounded in the same probability theory as the trust work upon which the model is loosely based. Furthermore, in its current capacity, the model does not guard against the rise of false but popular annotations (i.e., "folklore").

2 Incorporating a measure of credibility

2.1 A Motivating Example: Folklore and Popularity

To begin, we develop an example that exhibits the "folklore" problem in Champaign's model. We sketch what LOAR does with this and assume:

- A single annotator, a, who has created a set of 6 annotations/messages, $M_a = \{m_1, \ldots, m_6\}$. The 6^{th} annotation contains false information (e.g., claims cancer can be cured by magic crystals).
- Four peers, p_1 through p_4, who have experienced and rated a's annotations.
- A student s for whom we are trying to determine whether to recommend m_6 annotation.

Table 1 shows the ratings given by each respective participant to each $m_i \in M$. In Table 1, a's reputation is simply the mean average rating, T_q, which can be verified to be 0.6583. Next, we calculate the participant pairwise similary scores per Champaign's model (see Table 2). Note that we only show the relevant pairwise similarities (to student s)

Table 1. User message ratings

	m_1	m_2	m_3	m_4	m_5	m_6
p_1	0	1	1	1	0	1
p_2	1	1	1	1	0	1
p_3	0	0	0	1	0	0
p_4	0	1	1	0	1	1
s	1	1	1	1	1	?

Table 2. Peer similarities to student s

	p_1	p_2	p_3	p_4
s	0.2	0.6	-0.6	0.2

To calculate these scores, we simply find the proportion of common ratings (i.e., annotations that both parties have rated) that agree between both parties, viz., the number of times both parties rate an annotation the same (both 1 or both 0) divided by the total number of common ratings. This metric is a number in the interval $[0, 1]$, which we then map to the interval $[-1, 1]$ to derive the final similarity (to facilitate weighting votes). In particular, this allows the system to make recommendations based on consistently divergent behaviour (i.e., a peer whose annotation preferences are completely opposite the current student).

Next, LOAR uses the similarity scores along with peer scores for the marginal message in order to derive a trust metric. In particular, LOAR tallies the votes for and against a given message, and combines these tallies into a final metric (using, for example, the cauchy CDF). So, for example, p_1's vote for m_6 would increase the "votes for" tally by $1 + 1 * 0.2 = 1.2$. On the other hand, p_2 would increase the "votes for" tally by $1 + 1 * 0.6 = 1.6$, and $p3$ would increase the "votes against" tally by $1 + 1 * (-0.6) = 0.4$. Lastly, the system combines the ratings on m_6 with a's overall reputation to derive the predicted benefit, using either the Cauchy or Trust approach, which results in a trust metric of 0.86 and 0.76, for Cauchy and Trust, respectively.

Clearly then, m_6 has a very high predicted benefit. Accordingly, there is a high likelihood that m_6 will be shown to s (it will likely be over a predetermined threshold, and will be among the top annotations). However, it contains errors that could detract from learning, and so should in fact not be shown to s. This example shows how the popular opinion of a message could ultimately enable false information to spread in a network of peers.

2.2 New Trust Model to Address Folklore

At first glance, it appears as though the folklore problem could be addressed by classifying peers into different roles and weighting peer feedback according to these roles. For example, one could introduce two classes of users, "students" and "professors", and set professors' weights to infinity. This would allow professors to prevent false message propagation; a single bad vote from a professor would outweigh any number of votes from students. But even "professors" can be wrong, so instead of an infinite weight, one could set the weights according to some heuristic that allows for a sufficiently large number of other user roles to outweigh a "super user". Even then, it seems reasonable that the weights of users in any role should be variable, owing to the fact that users can make mistakes and can gain and lose credibility in a community. Accordingly, a simple static "weight" solution is insufficient and somewhat naive.

Instead, we proceed first by redefining the notion of "trust" in the online education scenario. In LOAR, a trust metric was annotation-specific, and corresponded to the predicted benefit of a given message, where predicted benefit was a combination of an annotator's mean reputation and the annotation's similarity-weighted rating. Instead of introducing a weight under the same model, we develop a new model drawing on the probability theory used throughout trust literature. We draw inspiration from PTM and TRAVOS in the use of Beta distributions, but continue to model the benefit of annotation objects themselves, drawing inspiration from BCM in this regard.

Recasting the trust model Determining whether an annotation or message will be well-received (i.e., is beneficial) is not deterministic; it can instead be modelled as a Bernoulli process. That is, if M is the event a message is well-received, then we seek to determine $\psi = Pr(M)$. Moreover, we allow this parameter to itself be represented as a random variable and rely on Bayes' theorem to update prior probability distributions over ψ. In particular, we can use the Beta distribution to represent the prior $Pr(\psi)$:

$$Pr(\psi) = Beta(\alpha^*, \beta^*) \tag{3}$$

However, since we model the trustworthiness of messages (not annotators), the user does not have any prior belief that directly corresponds to the message itself (he has yet to experience it, and so the only rational belief is to assume that $\alpha^* = \beta^* = 1$, i.e., that ψ is uniformly distributed in the interval $[0, 1]$). Accordingly, we construct a suitable belief by looking to the experiences of peers in the system, as in LOAR.

When a user solicits feedback about a message, his peers report binary ratings[3]. Equivalently, peers report parameters α_p and β_p such that $\alpha_p + \beta_p = 1$.

[3] For example, 0 could mean "did not find the message helpful" and 1 could represent "found the message helpful".

Initially, we restrict this report such that $\alpha_p, \beta_p \in \{0, 1\}$[4]. To combine peer reports, we model the similarity between users i and j using Hamming distance. The Hamming distance is a measure of the number of bits by which two binary strings differ, or equivalently, how many changes need to be made to string a to transform it into string b. Here, we can consider the series of common annotation ratings between two users to form "binary rating strings". (Table 1 above shows such a set of binary rating strings in the form of a matrix).

From the Hamming distance we derive the Hamming ratio between i and j, denoted hr_{ij} (the Hamming distance divided by the length of the binary strings, i.e., the number of common ratings). Since a Hamming distance of 0 means that the two strings are identical, a Hamming ratio of 0 suggests we simply take a peer report as given; in contrast, if the Hamming ratio is 1, we swap the values reported for α_p and β_p. We formalize this combination scheme as follows:

$$\alpha^* = 1 + \sum_{p \in P} (1 - hr_{sp}) \cdot \alpha_p + hr_{sp} \cdot \beta_p \tag{4}$$

$$\beta^* = 1 + \sum_{p \in P} (1 - hr_{sp}) \cdot \beta_p + hr_{sp} \cdot \alpha_p \tag{5}$$

Here, P is the set of all peers. This combination capitalizes on the fact that the Beta distribution is well-defined for all real-valued parameters $\alpha, \beta > 0$. Moreover, it allows us to easily extend peer reports to include expectations on message trust values. That is, a report $r \in [0, 1]$ can be translated into parameters $(\alpha, \beta) = (r, 1-r)$ so that a report of $r = 1$ corresponds to $\alpha = 1, \beta = 0$, a report of $r = 0.5$ corresponds to $\alpha = \beta = 0.5$, and $r = 0$ to $\alpha = 0, \beta = 1$. Thus, a user can solicit feedback from peers about an annotation even if those peers have yet to personally experience the annotation. This is useful if, for example, the current user has no or limited ratings in common with peers who have rated the annotation in focus. (That is, it might be more useful to use a report of expected usefulness from a peer who is highly similar to the current user rather than use an explicit report from a peer with whom the user has no history and thus no notion of similarity).

Incorporating a measure of credibility Under the new trust framework described above, we now introduce the notion of credibility. Credibility is a measure of the extent to which users should trust the opinions of peers within the community. Thus, credibility influences the similarity weighted Beta distribution derived above. In particular, we now also seek to determine $\kappa = Pr(C)$, where C is the event that a peer report is credible.

As before, we assume that κ is randomly distributed and can be described by a Beta distribution. Thus, the credibility metric $E(Pr(\kappa))$ is reported alongside peer ratings of annotations. For now, we assume that this credibility score is

[4] A report of 1 corresponds to the combination $(\alpha_p, \beta_p) = (1, 0)$ whereas a report of 0 corresponds to the combination $(\alpha_p, \beta_p) = (0, 1)$.

made available by an oracle, and we leave a discussion of one possible derivation to the next section[5].

The question that remains is how a user should combine his knowledge of peer credibility and a particular annotation's reputation gleaned through peer reports. When a user is highly similar to the peer from whom he receives a report, this combination is trivial; credibility can directly discount the reported rating. That is, one should listen to the advice of highly credible and similar peers more than the advice of non-credible peers. However, a difficulty arises when a user is dissimilar from a credible peer. In this circumstance, our above model will reverse the opinion of a credible peer. In some instances, this reversal could actually detract from the user's learning.

To make this more explicit, suppose that user i solicits advice from user j about a message m. Suppose further that the Hamming ratio between i and j is 1 (that is, they are completely opposite). Then, if j reports $(\alpha, \beta) = (0, 1)$ (i.e., he thinks the message not useful, or perhaps even incorrect), the similarity weighting scheme described above will reverse this opinion to $(\alpha, \beta) = (1, 0)$ when determining the trust metric from i's perspective. That is, the message, which j thinks should not be shown, will now be more likely to be shown. However, in this case, if j is perfectly credible, his opinion of a message corresponds to a very credible one. Accordingly, his report might be better taken verbatim rather than dampened by the Hamming ratio.

Accordingly, we propose a scheme detailed in Algorithm 1. This algorithm computes a trust metric by discounting peer reports by their community credibility, except when they report negatively on the given annotation. When this happens, the peer's negative rating is weighted using a combination of similarity and credibility. In particular, the role that similarity plays in blending the reported message rating is linearly reversed as the peer's credibility approaches 1 (i.e., perfect credibility). This credibility weighting scheme helps to address the issue of folklore propagation in an e-learning system. That is, highly credible peers (like professors and TAs) who report negatively about a given annotation will hold more sway than a number of less credible peers, even if the credible voters usual voting patterns tend to make them dissimilar to the current student.

Incorporating annotator reputation Lastly, we address the notion of an annotator's reputation. It is useful to model an annotation's reputation using some combination of explicit annotation ratings and the annotator's inherent reputation, especially when the given annotation has little or no explicit ratings. We will assume that an annotator's reputation is the same as his credibility (described and used above). In order to calculate the credibility for user u, we propose the following heuristic, inspired by the recursive derivation of credibility evidence variables in BCM: that a peer is considered credible if credible peers vouch for his annotations. We can derive such a credibility score as follows:

[5] In particular, we will use the following heuristic: users will be considered credible if many credible peers rate their messages as credible, a tenet of BCM [10]

Algorithm 1: Deriving A Trust Score Using Similarity and Credibility

```
Input: The current user, $u$, his set of peers, $P$, their credibility scores,
       $c_p \in [0,1]$, and their corresponding ratings for the annotation in
       focus, $r_p \in [0,1]$
Output: Parameters $\alpha^*$ and $\beta^*$ to a Beta distribution describing trust in
        the current annotation

$\alpha^* = \beta^* = 1$ // At the start, user has uniform expectation
foreach $p \in P$ do
    $hr_{up} \longleftarrow computeHammingRatio(u, p)$
    $(\alpha_p, \beta_p) \longleftarrow (r_p, 1 - r_p)$
    if $r_p == 0$ then
        // Adjust the similarity weight by credibility:
        $\alpha^* + = [1 - hr_{up} \cdot (1 - c)]\alpha_p + hr_{up}(1 - c) \cdot \beta_p$
        $\beta^* + = [1 - hr_{up} \cdot (1 - c)]\beta_p + hr_{up}(1 - c) \cdot \alpha_p$
    else
        // Else simply compute a credibility-dampened trust score
        $\alpha^* + = c_p \cdot [(1 - hr_{up})\alpha_p + hr_{up}\beta_p]$
        $\beta^* + = c_p \cdot [(1 - hr_{up})\beta_p + hr_{up}\alpha_p]$
    end
end
```

1. User credibility is given by $\kappa_u \sim Beta(\alpha_{c_u}, \beta_{c_u})$. In this paper, we assume that a single, global credibility distribution across all subjects suffices to describe the reputation of an annotator (versus a topic-specific metric).
2. When a peer p rates message m_u authored by u, p's report updates α_{c_u} and β_{c_u} as follows: a positive (negative) rating will increment α_{c_u} (β_{c_u}) by 1.
3. A rating by p should also only affect u's credibility to the extent that p is credible. That is, when p rates m_u, the κ_u hyperparameters are incremented as above, except that p's report is discounted by $E(\kappa_p)$. Thus, if p rates m positively and p is perfectly credible, i.e., $E(\kappa_p) = 1$, α_c is incremented by 1. If p is not credible at all, i.e., $E(\kappa_p) = 0$, then α_c is incremented by 0 (i.e., non-credible peers cannot influence u's credibility).

Ultimately, the annotation-specific reputation and annotator credibility can be combined to finalize a trust metric using either of the schemes proposed in LOAR (e.g., Cauchy-based combination)[6].

Example Revisited Returning to the example in §2.1, we present the LOAR similarities and Hamming ratios together in Table 3. Here we see very clearly the relationship between the similarity metric used by LOAR and the Hamming ratio. In particular, a higher Hamming ratio corresponds to a similarity that is closer to -1. A mapping $f : h \mapsto s$ from row h to row s is defined as $f(h) =$

[6] This derivation of credibility only works for those users who have created annotations and is only useful if those annotations have received ratings.

Table 3. Similarities and Hamming Ratios

	p_1	p_2	p_3	p_4
s	0.2	0.6	-0.6	0.2
h	0.4	0.2	0.8	0.4

Table 4. α and β reports ($c_i = 1$)

	p_1	p_2	p_3	p_4
h	0.4	0.2	0.8	0.4
(α_p, β_p)	(1,0)	(1,0)	(0,1)	(1,0)
(α'_p, β'_p)	(0.6,0.4)	(0.8,0.2)	(0,1)	(0.6,0.4)

Table 5. α and β reports ($c_i = 0$)

	p_1	p_2	p_3	p_4
h	0.4	0.2	0.8	0.4
(α_p, β_p)	(1,0)	(1,0)	(0,1)	(1,0)
(α'_p, β'_p)	(0,0)	(0,0)	(0.8,0.2)	(0,0)

$1 - 2 \cdot h$. Hence, these metrics fundamentally measure the same thing and differ only by an affine transformation.

The cred-trust algorithm dampens trust values according to peer credibility. A peer is credible if credible peers rate their annotations highly, or if credible peers rely on their ratings. To begin, let us assume that all peers are perfectly credible, i.e., that $c_i = 1$. Table 4 shows the initial α_p, β_p reports as well as their credibility-weighted values α'_p, β'_p, as given by Algorithm 1.

Using these values, we can see that the reported trust metric would be 0.5. Upon reflection, this trust value makes sense. Similarity forms a continuum between "exactly like me" and "exactly opposite me". Peers p_1 and p_4 have Hamming ratios of 0.4, indicating they are centered in that continuum. Accordingly, one cannot gain much insight from their reports. Furthermore, p_2 (p_3) has a Hamming ratio of 0.2 (0.8). On the balance, these two peers' opinions should offset each other. Ultimately, we cannot learn anything about the trustworthiness of the medium in this case, since all peers are equally and perfectly credible.

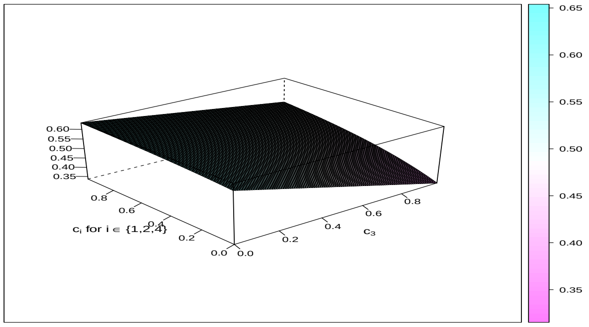
Table 5 illustrates the case where all peers have a credibility score of 0. These combinations result in a trust metric equal to 0.6. In this case, the algorithm completely disregards the votes of peers who liked the given message. However, p_3, who is almost completely opposite to s, disliked the message. Even though p_3 is not credible at all, or perhaps because he is not credible, the algorithm reverses his opinion, resulting in a trust metric of 0.6.

As a final example, suppose we let $c_1 = c_2 = c_4 = c$, $c \in [0, 1)$, and $c_3 = 1$. The interplay between credibilities for the two peer sets is shown in Figure 1. In the cases where p_1, p_2, p_3 have bad trust, their opinions will be discounted, while p_3's opinion will be taken verbatim.

3 Conclusions

In summary, we have developed a new trust model for use in determining the reputability of a particular message in a social networking environment, motivated by earlier work on limiting the presentation of annotations in repositories of learning objects, for peer-based intelligent tutoring. Our model incorporates feedback from peers and then weights this feedback by Hamming similarity. The combination of peer feedback is accomplished using a Beta distribution, and it

Fig. 1. Credibility impact on trust under the current ratings profile



takes advantage of various properties of the distribution (notably, that it is well-defined for real-valued hyperparamters, as in BRS). However, the model also combines feedback according to peer credibility, inspired by the use of that term in BCM. This distinguishes our approach from trust models that only discount reputability based on similarity (as in PTM). The use of an external measure of credibility in addition to similarity allows the model to combat the spread of false information in the system.

Our work contrasts with that of other researchers. The Bayesian learning trust model BLADE [9] for modeling seller/advisor trustworthiness employs Dirichlet distributions to allow multiple dimensions to be considered. Additionally, BLADE tries to learn the evaluation function of agents and thus does not simply ignore reports deemed untrustworthy. While BLADE does address subjective differences, this notion of similarity is not also combined with a modeling of credibility and thus our approach goes beyond BLADE's focus, which may be best seen as a kind of reputation alignment [6]. Other researchers exploring trust in social networks are focused more on how to propagate opinions amongst peers. For example, Hang, Zhang and Singh [3, 4] suggest which peers might have the most valuable advice to offer (with research that evaluates the trustworthiness of a witness in terms of trust it puts in common acquaintances). That work, however, does shed light on peer credibility and as such may be of value for us to explore for future extensions of our model, which instead focuses on which messages to recommend to users. In a similar vein, the concept of confidence explored in [7] or the suggestion of measuring credibility in terms of behavioural models [8] may also serve as starting points for deepening our solution.

As a final remark on this model, it would be beneficial to incorporate confidence thresholds into peer reports. In so doing, peers who have extensive evidence about a message's expected usefulness (either through a number of ratings-in-common with other users who rated the current message or via numbers of peer interactions with the given annotator) could report more confidently. This idea, inspired by TRAVOS, is left as a future extension. Of interest for future work as well is to do a head to head comparison with the LOAR model, in order to quantify the benefits accrued from our model in correcting for the potential of folklore and in expanding peer reports to include expectations about message benefits. Another important avenue is to explore learning to determine a more precise relationship between rating Hamming distances and the probability a user will find an annotation useful. A final open direction is to return to Gorner's question [2]: how best to determine the size and composition of a social network, considering whether the number of peers whose advise is sought should be unbounded as in LOAR or delimited as in PTM or BCM.

References

1. John Champaign, Jie Zhang, and Robin Cohen, *Coping with poor advice from peers in peer-based intelligent tutoring: The case of avoiding bad annotations of learning objects*, Proceedings of User Modeling, Adaptation and Personalization (UMAP), 2011, pp. 38–49.
2. Joshua Gorner, Jie Zhang, and Robin Cohen, *Improving trust modelling through the limit of advisor network size and use of referrals*, Electronic Commerce Research and Applications (ECRA) (2012), accepted.
3. Chung-Wei Hang and Munindar P. Singh, *Generalized framework for personalized recommendations in agent networks*, Autonomous Agents and Multi-Agent Systems (AAMAS) (2011), 1–27.
4. Chung-Wei Hang, Zhe Zhang, and Munindar P. Singh, *Generalized trust propagation with limited evidence*, IEEE Computer (2012), 1–8.
5. Audun Jøsang and Roslan Ismail, *The beta reputation system*, Proceedings of the 15th Bled Electronic Commerce Conference, 2002, pp. 324–337.
6. Andrew Koster, Jordi Sabater-Mir, and Marco Schorlemmer, *Inductively generated trust alignments based on shared interactions*, Proceedings of the 9th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 2010, pp. 1571–1572.
7. Ugur Kuter and Jennifer Golbeck, *Sunny: A new algorithm for trust inference in social networks, using probabilistic confidence models*, Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07), 2007, pp. 1377–1382.
8. Zeinab Noorian, Stephen Marsh, and Michael Fleming, *Multi-layer cognitive filtering by behavioral modeling*, Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2011, pp. 871–878.
9. Kevin Regan, Pascal Poupart, and Robin Cohen, *Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change*, Proceedings of the 21st National Conference on Artificial Intelligence (AAAI), 2006, pp. 1206–1212.
10. A. Seth, J. Zhang, and R. Cohen, *Bayesian credibility modeling for personalized recommendation in participatory media*, Proceedings of the International Conference on User Modeling, Adaptation and Personalization (UMAP), 2010, pp. 279–290.
11. W. T. Luke Teacy, Jigar Patel, Nicholas R. Jennings, and Michael Luck, *Travos: Trust and reputation in the context of inaccurate information sources*, Autonomous Agents and Multi-Agent Systems (AAMAS) **12** (2006), no. 2, 183–198.
12. Jie Zhang and Robin Cohen, *Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach*, Electronic Commerce Research and Applications (2008), 330–340.