Controversial Users demand Local Trust Metrics: an Experimental Study on Epinions.com Community

Paolo Massa and **Paolo Avesani**
ITC-iRST
Via Sommarive 14 - I-38050 Povo (TN) - Italy
{massa,avesani}@itc.it

Abstract

In today's connected world it is possible and very common to interact with unknown people, whose reliability is unknown. Trust Metrics are a recently proposed technique for answering questions such as "Should I trust this user?". However, most of the current research assumes that every user has a global quality score and that the goal of the technique is just to predict this correct value. We show, on data from a real and large user community, *Epinions.com*, that such an assumption is not realistic because there is a significant portion of what we call *controversial users*, users who are trusted and distrusted by many. A global agreement about the trustworthiness value of these users cannot exist. We argue, using computational experiments, that the existence of controversial users (a normal phenomena in societies) demands Local Trust Metrics, techniques able to predict the trustworthiness of an user in a personalized way, depending on the very personal view of the judging user.

Introduction

In today's connected world, it is possible and common to interact with unknown people. This happens when contacting a stranger via her email address found on the Web, using a site that allows messaging between users or reading, on an opinions site, a review of a product written by someone we don't know.

In this *uncertain* world, it is necessary to take into account questions such as "Should I trust this person?". The emerging way of dealing with this new requirement is to allow all the users to express their level of trust on other users, aggregate this information and reason about it. This intuition is exploited in modern search engines such as *Google.com*, that considers a link from one site to another as an expression of trust (Page *et al.* 1998), in e-marketplaces such as *Ebay.com*, that allows users to express their level of satisfaction after every interaction with another user and has been suggested for peer-to-peer systems where peers keep a history of interactions with other peers and share this information with the other peers (Cornelli *et al.* 2002).

A considerable amount of research has been carried on recently on these and related topics, such as Reputation Systems (Resnick *et al.* 2000), Trust Metrics (Golbeck, Hendler, & Parsia 2003; Ziegler & Lausen 2004; Levien 2003; Massa & Avesani 2004; Guha *et al.* 2004) and personalizing PageRank (Haveliwala, Kamvar, & Jeh 2003).

However most of the current research takes the assumption that every user[1] has an objective trustworthiness value and the goal of the techniques is just to guess this correct value. Conversely, we think that such an assumption is misleading. We argue that these techniques should take into account the fact that different users can have different opinions about a specific user.

Hence we distinguish between Global and Local Trust Metrics (Massa & Avesani 2004; Ziegler & Lausen 2004). Both try to predict the trustworthiness[2] of a given user. Global Trust Metrics assign to a given user a unique trust score, the same independently of the user that is evaluating the other user's trustworthiness. On the other hand, a Local Trust Metric provides a personalized trust score that depends on the point of view of the evaluating user.

In this paper, we will devote special attention to *controversial users*. A controversial user is a user that is judged by other users in very diverse ways, for example, she is trusted or appreciated by many and is distrusted or negatively rated by many.

Hence, the goal of the paper is to investigate the differences between Global and Local Trust Metrics, concentrating on controversial users and considering situations where one technique is more appropriated than the other. Data from the large *Epinions.com* community confirm our hypothesis that in complex soci-

[1] From now on, we will use the term "user" in order to indicate an autonomous entity able to express and receive trust statements. However the same concepts apply also if considering peers in a peer-to-peer system, interacting servers, agents in a multi agents system or communicating mobile devices such as mobiles and robots.

[2] Different authors use different terms. Other used terms are authority, reputation and reliability. The terms often represent the same concept but are used in slightly different contexts. We use trust and trustworthiness.

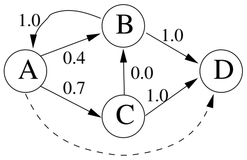

Figure 1: Trust network. Nodes are users and edges are (trust) statements. The dotted edge is one of the undefined and predictable trust statements.

eties there is a non negligible percentage of controversial users. The experiments we conducted show that a Local Trust Metric achieves higher accuracy than a global one in predicting the trust a specific user should place into a controversial user.

The rest of the paper is structured as follows. First, we provide definitions for concepts such as trust and trust network. Then we explain trust metrics focusing especially on the differences between local and global ones and introduce the ones we chose for the experiments. Finally we present the experiments we carried out and discuss the results.

Trust Networks and Trust Metrics

We call *trust statement* the explicit opinion expressed by an user about another user regarding the perceived quality of a certain characteristic of this user. For example, on a site where users contribute reviews about products, users could be asked to express a positive trust statement in a user "whose reviews and ratings they have consistently found to be valuable"[3]. We model trust as a real value in the interval $[0, 1]$, where $T(A, B) = 0$ means that A has issued a statement expressing that her degree of trust in B is the minimum, i.e. she distrusts totally B. On the other hand, $T(C, B) = 1$ means that C totally trusts B. As the previous examples show, trust statements are subjective: a user can receive different trust scores from different users. In most settings, a user has a direct opinion only about a very small portion of users. The remaining users are unknown users. By aggregating all the trust statements expressed by every user, we can produce the global trust network (or social network). An example of a simple trust network can be seen in Figure 1. As a consequence of the previously introduced properties of trust, such network is a directed, weighted graph whose nodes are users and whose edges are trust statements.

Trust Metrics (Golbeck, Hendler, & Parsia 2003; Levien 2003; Ziegler & Lausen 2004; Massa & Avesani 2004) are a technique used to predict trust scores of users. Given a current user A, they try to predict the trust score of the users unknown to A, by exploiting controlled trust propagation. For example, in the social network of Figure 1, a trust metric can be used to predict the trust A could place in D. The common assumption exploited in trust metrics is that if user A trusts B at a certain level and B trusts C at another level, something can be inferred about the level of trust of A in C.

Trust metrics can be classified into global and local ones (Massa & Avesani 2004; Ziegler & Lausen 2004). Local trust metrics take into account the subjective opinions of the active user when predicting the trust she places in unknown users. For this reason, the trust score of a certain user can be different when predicted from the point of view of different users. Instead, global trust metrics compute a trust score that approximates how much the community as a whole trusts a specific user. The formal definition of a global trust metric is hence $T : U \to [0, 1]$ while local trust metrics are defined as $T : U \times U \to [0, 1]$.

PageRank (Page *et al.* 1998), one of the algorithm behind the search engine *Google.com*, is an example of global metric since the computed authority of a certain Web page is the same for every user independently of her sites preferences. In general, while local trust metrics can be more precise and tailored to the single user's views, they are also computationally more expensive, since they must be computed for each user whereas global ones are just run once for all the community. Another interesting feature of local trust metrics is the fact they can be attack-resistant (Levien 2003): users who are considered malicious (from a certain user's point of view) are excluded from trust propagation and they don't influence the personalization of users who don't trust them explicitly. (Gori & Witten 2005) shows that malicious exploitation of links is an inherent and unavoidable problem for global trust metrics. The rise of link-farms, that provide links to a site in order to increase its PageRank, also makes evident the problem.

The differences between local and global trust metrics are especially evident when considering *controversial users.* We will define more precisely controversial users in next sections, however it should be clear that global trust metrics are not suited for this type of users since an *average* trust score on which all the users might agree does not exist.

A Local Trust Metric

In this section we introduce the local trust metric we used in our experiments. We chose to use the MoleTrust trust metric, introduced in (Massa, Avesani, & Tiella 2005). The choice was guided by the need of a time-efficient local trust metric, since the number of trust scores to be predicted in the experiments is very large. MoleTrust predicts the trust score of *source user* on *target user* by walking the social network starting from the source user and by propagating trust along trust edges. Intuitively the trust score of a user depends on the trust statements of other users on her and their trust scores. The pseudocode is presented in Figure 2. Precisely, the MoleTrust trust metric can be modeled in 2 steps. The purpose of the first step is to destroy cycles in the graph. The problem created by cycles is

[3] From the *Epinions.com* Web of Trust FAQ (http://www.epinions.com/help/faq/?show=faq_wot)

that they require visiting a node many times adjusting progressively the temporary trust value until this value converges. In order to have a time-efficient algorithm, it is preferable to visit every user just once and, in doing this, to compute her definitive trust value. In this way, the running time is linear with the number of nodes. So the first step modifies the social network by ordering users based on distance from source user and keeping only trust edges that goes from users at distance n to users at distance $n + 1$. The *trust propagation horizon* specifies the maximum distance from source user to which trust is propagated along trust chains. This reduces the number of visited users and hence achieves shorter computational time.

```
Step 1:
Input:$source\_user$, $trust\_net$, $trust\_prop\_horizon$
$dist = 0$; $users[dist] = source\_user$
while ($dist \leq trust\_prop\_horizon$) do
  $dist++$
  $users[dist]$=users reachable from $users[dist-1]$
    and not yet visited
  keep $edges$ from $users[dist-1]$ to $users[dist]$

Step 2:
Output: $trust\_scores$ for users
$dist = 0$; $trust(source\_user) = 1$
while ($dist \leq trust\_prop\_horizon$) do
  $dist++$
  foreach $u$ in $users[dist]$
    $trust(u) = \frac{\sum_{i=pred(u)}(trust(i)*edge(i,u))}{\sum_{i=pred(u)}(trust(i))}$
```

Figure 2: MoleTrust pseudocode. $pred(u)$ returns predecessors p of user u for which $trust(p) \geq 0.6$. $edge(i, u)$ is the value of the statement issued by i on u.

After step 1, the modified social network is a reduced directed acyclic graph, with trust flowing away from source user and never flowing back.

The second step is a simple graph walk over the modified social network, starting from source user. The trust score of one user at distance x only depends on trust scores of users at distance $x - 1$, that are already computed and definitive. For predicting the trust score of a user, MoleTrust analyzes the incoming trust edges and discards the ones coming from users with a predicted trust score less than 0.6. These users are not trustworthy and their trust statements are ignored: this avoids situations in which the trust score of an unknown user depends only on statements issued by untrustworthy users. The predicted trust score of a user is the average of all the accepted incoming trust edge values, weighted by the trust score of the user who has issued the trust statement. The MoleTrust trust metric is able to compute trust values only in users reachable from the source user and inside the trust propagation horizon.

Experiments on Epinions.com

We conducted the experiments on data of the community of users of the popular Web site *Epinions.com*. *Epinions.com* is a web site where users can write reviews about products and assign them a rating. *Epinions.com* also allows the users to express their *Web of Trust*, i.e. "reviewers whose reviews and ratings they have consistently found to be valuable" and their *Block list*, i.e. a list of authors whose reviews they find consistently offensive, inaccurate, or in general not valuable. Inserting a user in the Web of Trust equals to issuing a trust statement in her ($T(A, B) = 1$) while inserting her in the Block List equals to issuing a distrust statement in her ($T(A, B) = 0$). Intermediate values such as 0.7 are not expressible on *Epinions.com* and hence not available in our experiments.

The *Epinions.com* dataset we used contained $\sim$132000 users, who issued $\sim$841000 statements ($\sim$717000 trusts and $\sim$124000 distrusts). $\sim$85000 users received at least one statement. Based on the actual characteristics of available data, particularly the fact that statements values are just 1 and 0 and not any real in the interval $[0, 1]$, we now define some quantities. The *controversiality level* of a user is the number of users who disagree with the majority in issuing a statement about that user. For example, a user who received 21 distrust statements and 14 trust statements has a controversiality level of 14. Formally, $controversiality\_level = min(\#trust, \#distrust)$. A user who has a controversiality level of x is called *x-controversial*. *0-controversial users* received only trust or distrust statements and they are non controversial. We define a user who has a controversiality level not less than x as a *at least x-controversial* user. As one might expect, most of the users are non controversial, in the sense that all the users judging them share the same opinion. Out of the 84601 users who received at least one statement, 67511 are 0-controversial, 17090 (more than 20%) are at least 1-controversial, i.e. at least one user disagrees with the others, 1247 are at least 10-controversial, 144 are at least 40-controversial and one user is 212-controversial.

However, a user with 100 trusts and 5 distrusts and a user with 5 trusts and 5 distrusts have the same controversiality level, even if the first one is much less controversial. For this reason, we define another quantity, *controversiality percentage*, as $\frac{\#trust-\#distrust}{\#trust+\#distrust}$. A user with 1 ($-1$) as controversiality percentage is trusted (distrusted) by all her judgers. A user whose controversiality percentage is 0 is highly controversial since other users split into 2 opinions groups of same size.

Let us now specify the global trust metric we used in our experiments. The choice was guided by the fact that statements are binary (1 and 0) and not continuous. So, we chose to use a very simple metric that is similar to the one used by the online auctions site *Ebay.com*. In order to predict the trust score of one user, the global trust metric (in the following called *ebay*) simply computes the fraction of received trust statements over all the received statements. Formally, $trust_{ebay}(user) = \frac{\#trust}{\#trust+\#distrust}$. The trust score of a user with 0 received statements is not predictable.

We considered using more complex global metrics but none of them seemed suited for the available data. In particular, PageRank (Page *et al.* 1998), probably the most advanced global metric, is not suited for the task since the original formulation does not take into account the concept of negative links and also does not produce an "authority" value in the interval $[0, 1]$; adapting it would have meant changing it profoundly and introducing additional biases and noises. It should be noted that, even if the input trust statement values are just 1 and 0, both the local and global metric predict trust scores in the interval $[0, 1]$.

The evaluation technique is a standard one in machine learning: leave-one-out. Taken one trust statement from user A to user B, we remove it from the trust network and try then to predict it using the local trust metric. We then compare the predicted trust score against the original trust statement. For the global trust metric, we compare the predicted global trust score of B against the statement issued by A on B. Two measures are derived from this evaluation technique: accuracy and coverage. Accuracy represents the error produced when predicting a score. We use Mean Absolute Error that consists in computing the absolute value of the difference between the real score and the predicted one. Coverage refers to the ability of the algorithms to provide a prediction. In this case we compute the percentage of predictable trust statements.

Results

Figure 3 and Figure 4 shows the prediction errors for *moletrust2*[4] and *ebay*, respectively. The x axis represents the controversiality level of users. The plotted value is the mean absolute error over the statements received by users who are at least x-controversial. We distinguish the error over only trust statements (bottom line), over only distrust statements (top line) and over all the statements (central line). The graphs show how predicting distrust statements is more difficult. However while for ebay the error on distrust statements is higher than 0.6, for moletrust2 it is around 0.4.

On the other hand, the error over trust statements is very similar for the 2 different techniques. This is because the number of trust statements is much larger than the number of distrust statements: ∼717000 against ∼124000. This fact does not allow to clearly distinguish how much trust metrics are effective: a technique predicting almost always a trust score close to 1 (as ebay does) produces a very small error over trust statements.

So we concentrate on *controversiality percentage*. Figure 5 and Figure 6 shows the accuracy of moletrust2 and ebay over buckets grouping users with same controversiality percentage. While the mean absolute error close to the borders (1 and -1 that represents users who are non controversial) is similar for the 2 algorithms, moletrust2 is able to significantly reduce the error for controversial users, users close to the centre of the graph. As expected, the error produced by ebay on users with controversiality percentage 0 is 0.5 since these users received n trusts and n distrusts and the metric predicts 0.5 as the global trust score for them and, when compared to the real trust score (either 1 for trust or 0 for distrust), encounters an error of 0.5 in every single case. This is an inherent limit for global trust metrics that on users with controversiality percentage of 0 cannot achieve an error smaller than 0.5.

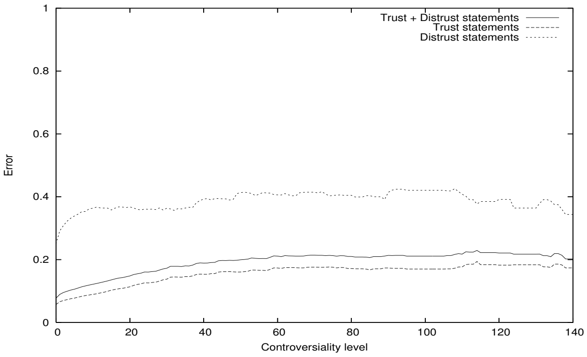

Figure 3: Moletrust2 prediction error on different kinds of statements (trust, distrust and both) for users who are *at least x-controversial.*

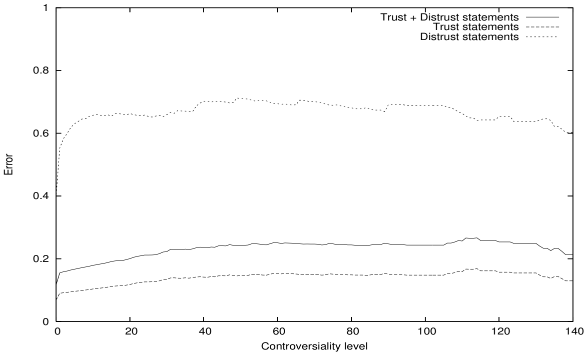

Figure 4: Ebay prediction error on different kinds of statements (trust, distrust and both) for users who are *at least x-controversial.*

It should be noted, however, that the majority of users fall into the buckets near the borders. The number of statements is ∼841000. Of them, ∼440000 (more than 50%) go into users who are in the 1 bucket and ∼206000 for the 0.9 bucket. ∼41000 go into the -1 bucket. Only 1972 statements go into users falling into the 0 bucket and 1013 into the -0.1 bucket, the least populated. Once more, this data shows that most of the users are not (or little) controversial. However, the

[4] *Moletrust2* represents the local trust metric with trust propagation horizon set to 2.

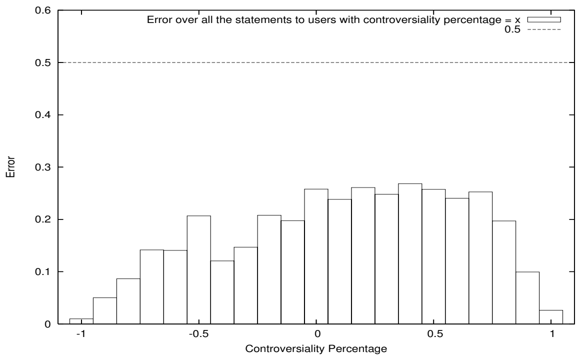

Figure 5: Moletrust2 prediction error for users grouped by controversiality percentage. Users at the borders are non controversial while users at the centre are highly controversial.

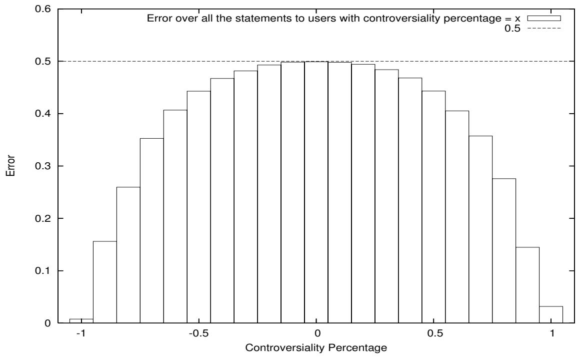

Figure 6: Ebay prediction error for users grouped by controversiality percentage.

fraction of controversial users is not negligible and it can be argued that these users are the ones on which a trust metric is more required. Figure 5 and Figure 6 show how, for controversial users (located around bucket 0), our local trust metric significantly reduces error when compared to the global trust metric.

It is interesting to separate the error generated in predicting trust and distrust statements (Figure 7). As expected, for users close to -1 bucket (distrusted by almost all their judgers), it is easy to correctly predict (the many) distrust statements and it is hard to predict (the few) trust statements. The opposite is true for users with controversiality percentage close to 1. If we compare the error over distrust statements for moletrust2 and ebay, we observe how much the local metric is able to reduce the error. A similar observation can be made for trust statements as well. However, it is not clear why the error for moletrust2 on distrust statements is smaller than the one for trust statements also for users who are mostly trusted by others, such as users in the 0.4 bucket. Since most of the users (more than 75%) fall into bucket 1 and 0.9, these results are not in contrast with the previous ones showing that the error is greater on distrust than trust statements.

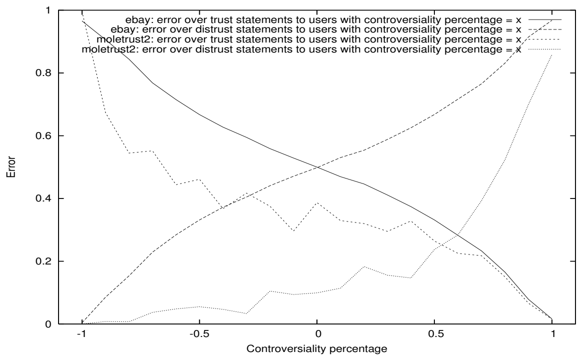

Figure 7: Prediction error of moletrust2 and ebay, considered separately for trust and distrust statements.

Another evaluation quantity, besides accuracy, is coverage, i.e. the percentage of statements that are predictable by the algorithms. The coverage of a global trust metric is very close to 1, since a single statements received by a user is enough for predicting a trust score for that user. Instead, the chosen local trust metric is able to predict a trust score only if there is at least one trust path from source user to target user, shorter than the trust propagation horizon.

For reasons of space, we don't insert the graphs that show the coverage of the different algorithms for different controversiality levels. However, they don't present surprising properties. The coverage of moletrust2 is around 0.8 for distrust statements, 0.88 for trust statements and 0.86 on both. These percentages are stable across all the controversiality levels.

Running Moletrust with different trust propagation horizons (2, 3 and 4) does not produce very different results. The accuracy is very similar for the 3 configurations across all the levels and percentage of controversiality. The coverage is, as expected, higher for moletrust4 that for moletrust2 since it is possible to propagate further the trust and hence to reach more users, who might have expressed a statement about the target user. However, a larger trust propagation horizon also means greater computational time and this can be an issue if the user need to have a result in real time. In this paper we concentrate more on the differences between one local trust metric (moletrust2) and one global trust metric (ebay) and so we don't analyze the differences produced on MoleTrust performances by different trust propagation horizon values. As already stated, these differences are not very significant, at least with respect to the accuracy.

Discussion of Results

Since Moletrust is a local trust metric, it provides personalized trust scores that depends on the judging user. For this reason, it is expected to work better than a global one, especially when predicting the trust score of controversial users that, by definition, don't have a *global* trustworthiness value agreed by all the users.

It is worth recalling that on *Epinions.com community*, controversial users are a non negligible fraction: 20% of the users who received a statement are at least 1-controversial.

Figure 6 shows that global metric error for highly controversial users is, as expected, 0.5. However, Moletrust does not perform as well as one could have desired. One reason could be that the *Epinions.com* dataset we used is very sparse and trust propagation often cannot reach a user from many trust paths. It remains an open point to verify if, on a more connected community, a local trust metric can decrease the error close to 0 also for controversial users.

The experiments clearly show that correctly predicting a distrust statement is harder than predicting a trust statement. However, it is very important to inform the user about other users who should not be trusted, such as a malicious user that is trying to fool the active user for her personal benefit: for example by rating highly her own book on an opinions site. Correctly predicting distrust is hence an important research challenge. We believe this should be carried on without assuming a global measure of “goodness” or “badness” but that users have a subjective notion of who to trust or not.

A global trust metric can be run once for the all community, while a local one must be run once for every single user, in order to predict the trust scores of other users from her personal point of view. This fact makes local trust metrics difficult to integrate into a centralized service such as *Google.com* because of computational time and storage problems. The more reasonable setting for a local trust metric is the one in which every user runs it from her personal point of view, possibly in her mobile device or in her browser.

Another weak point of local trust metrics is the reduced coverage: while global trust metrics coverage is usually close to 1, this is not always the case for local ones. We have seen how on the *Epinions.com* community, moletrust2 is able to reach on average a good coverage (almost 0.8). A possible improvement would be to integrate the 2 techniques, for example, by using a global metric when the local one fails.

Eventually, in non controversial domains, global metrics can be more suited because they guarantee greater coverage, smaller computational time with similar accuracy. For example, on *Ebay.com*, the notion of good seller is a shared concept agreed by most of the users. Maybe some users give more importance in timeliness in sending the goods while others care more about correspondence between photo and shipped good but what makes a good seller is an unambiguous concept. When we move into more subjective domains, such as evaluating music tracks or movies (or even discussing political ideas), it is reasonable to accept significant disagreement between users. In this contexts, a local trust metric can be more effective. However in both cases, if there are controversial users, a local trust metric is probably more suited.

Conclusions

In this paper we analyzed the differences in accuracy and coverage of local and global Trust Metrics. We especially concentrated on controversial users, defined as users that are judged in very different ways by other users. We have argued that controversial users are a non negligible portion of the users on the large *Epinions.com* community. We have introduced a local Trust Metric and compared it against a global trust metric in the task of predicting trust scores of unknown users. The results demonstrates that our local Trust Metric is able to significantly reduce the prediction error for controversial users, while retaining a good coverage.

Future works will involve comparing more Trust Metrics (both local and global ones) and also analyzing different communities of users, such as *Ebay.com*.

ACKNOWLEDGMENTS

The authors would like to thank *Epinions.com* for kindly making available the data for this study. We would in particular like to thank also R. Guha, Ravi Kumar, Prabhakar Raghavan and Andrew Tomkins.

References

Cornelli, F.; Damiani, E.; di Vimercati, S. D. C.; Paraboschi, S.; and Samarati, P. 2002. Implementing a reputation-aware gnutella servent. In *International Workshop on Peer-to-Peer Computing*.

Golbeck, J.; Hendler, J.; and Parsia, B. 2003. Trust networks on the Semantic Web. In *Proceedings of Cooperative Intelligent Agents*.

Gori, M., and Witten, I. 2005. The bubble of web visibility. *Commun. ACM* 48(3):115–117.

Guha, R.; Kumar, R.; Raghavan, P.; and Tomkins, A. 2004. Propagation of trust and distrust. In *Proceeding of WWW '04 conference*, 403–412. ACM Press.

Haveliwala, T.; Kamvar, S.; and Jeh, G. 2003. An analytical comparison of approaches to personalizing pagerank. Technical report, Stanford, USA.

Levien, R. 2003. *Advogato Trust Metric.* Ph.D. Dissertation, UC Berkeley, USA.

Massa, P., and Avesani, P. 2004. Trust-aware collaborative filtering for recommender systems. In *Proc. of Cooperative Information Systems Conf. (CoopIS)*.

Massa, P.; Avesani, P.; and Tiella, R. 2005. A Trust-enhanced Recommender System application: Moleskiing. In *Procedings of ACM SAC TRECK Track*.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1998. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford, USA.

Resnick, P.; Zeckhauser, R.; Friedman, E.; and Kuwabara, K. 2000. Reputation Systems. *Communication of the ACM* 43(12).

Ziegler, C., and Lausen, G. 2004. Spreading activation models for trust propagation. In *IEEE Conf. on e-Technology, e-Commerce, and e-Service (EEE'04)*.