Trustlet, Open Research on Trust Metrics.

Paolo Massa
Galliera Hospital
96 PUBLICATIONS **4,089** CITATIONS

Trustlet, Open Research on Trust Metrics

Paolo Massa and Kasper Souren

FBK/rst
Via Sommarive, 14
Povo (TN) - Italy
{massa,souren}@fbk.eu

Abstract. A trust metric is a technique for predicting how much a user of a social network might trust another user. This is especially beneficial in situations where most users are unknown to each other such as online communities. We think the recent tumultuous evolution of social networking demands for a collective research effort. With this in mind we created Trustlet.org, a platform consisting of a wiki for open research on trust metrics. The goal of Trustlet is to collect and distribute trust network datasets and trust metrics code as free software, in order to facilitate the comparison of different trust metrics algorithms and a more coherent progress in this field. At present we made available some social network datasets and code for some trust metrics. In this paper we also report a first empirical evaluation of different trust metrics on the Advogato social network dataset.

Key words: Trust Metrics, Social network analysis, Wiki, Advogato, Free software, Data acquisition, Science Commons

1 Introduction

In our current society it is more and more common to interact with strangers, people who are totally unknown to us. This happens for example when receiving an email asking for collaboration or advise from an unknown person, when we rely on reviews written by unknown people on sites such as Amazon.com, and also when browsing random profiles on social networking sites such as Facebook.com or Linkedin.com. Even more surprising is the fact a huge amount of commercial exchanges happen now between strangers, facilitated by platforms such as Ebay.com. In all systems in which it is possible to interact with unknown people, it is important to have tools able to suggest which other users can be trustworthy enough for engaging with. Trust metrics and reputation systems [1] have precisely this goal and become even more important, for instance, in systems where people are connected in the physical world such as carpooling systems or hospitality exchange networks (i.e. couchsurfing.com), in which users accept to have strangers into their car or their house.

A commonly cited definition of trust was proposed by Diego Gambetta: "Trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent will perform a particular action, both before [we] can monitor each action (or independently of his capacity of ever be able to monitor it) and in a context in which it affects [our] own action" [3]. In all the previous example it is possible to consider the social relationship users can express as a trust statement, an explicit statement stating "I trust this person in this context" (for example as a pleasant guest in a house or as a reliable seller of items) [2].

While research about trust issues spanned disciplines as diverse as economics, psychology, sociology, anthropology and political science for centuries, it is only recently that the widespread availability of modern communication technologies facilitated empirical research on large social networks, since it is now possible to collect real world datasets and analyze them [2]. As a consequence, recently computer scientists and physicists started contributing to this research field as well [4, 5].

Moreover we all start relying more and more on these social networks, for friendship, buying, working, ... As this field become more and more crucial, in the past few years many trust metrics have been proposed but there is a lack of comparisons and analysis of different trust metrics in the same conditions. As Sierra and Sabater put it in their complete "Review on Computational Trust and Reputation Models" [6]: "Finally, analyzing the models presented in this article we found that there is a complete absence of test-beds and frameworks to evaluate and compare the models under a set of representative and common conditions. This situation is quite confusing, specially for the possible users of these trust and reputation models. It is thus urgent to define a set of test-beds that allow the research community to establish comparisons in a similar way to what happens in other areas (e.g. machine learning)" (emphasis added). Our goal is to fill this void and for this reason we set up Trustlet [7], a collaborative wiki in which we hope to aggregate researchers interested in trust and reputation and build together a lively test-bed and community for trust metrics evaluation. A project with similar goals is the Agent Reputation and Trust (ART) Testbed [8]. However ART is more focused on evaluating different strategies for interactions in societies in which there is competition and the goal is to perform more successfully than other players, in a specific context. Our take with Trustlet is about evaluating performances of trust metrics in their ability to predict how much a user could trust another user, in every context. For this reason, we want also to support off-line evaluation of different trust metrics on social network datasets. The two testbeds are hence complementary.

In this paper we describe Trustlet, the reason behind its creation and its goals, we report the datasets we have collected and released and the trust metrics we have implemented and we present a first empirical evaluation of different trust metrics on the Advogato dataset.

2 Trust metrics

Trust metrics are a way to measure trust one entity could place in another. After a transaction user Alice on Ebay can explicitly express her subjective level of trust in user Bob. We model this as a trust statement from Alice to Bob. Trust statements can be weighted, for example on Advogato [9] a user can certify another user as Master, Journeyer, Apprentice or Observer, based on the perceived level of involvement in the free software community. Trust statements are directed and not necessary symmetric: it's possible a user reciprocates with a different trust statement or simply not at all. By aggregating the trust statements expressed by all the members of the community it is possible to build the entire trust network (see for example Figure 1). A trust network is hence a directed, weighted graph. In fact trust can be considered as one of the possible social relationships between humans, and trust networks a subclass of social networks [4, 5].

Trust metrics are tools for predicting the trust a user could have in another user, by analyzing the trust network and assuming that trust can somehow be propagated. One of the assumptions is that people are more likely to trust a friend of a friend than a random stranger [12, 10, 11, 9].

Trust metrics can either be local or global [10, 12]. A global trust metric is a trust metric where predicted trust values for nodes are not personalized. On the other hand, with local trust metrics, the trust values a user sees for other users depend on her position in the network. In fact, a local trust metric predicts trust scores that are personalized from the point of view of every single user. For example a local trust metric might predict "Alice should trust Carol as 0.9" and "Bob should trust Carol as 0.1", or more formally trust(A,C)=0.9 and trust(B,C)=0.1. Instead for global trust metrics, trust(A,B)=reputation(B) for every user A. This global value is sometimes called reputation [2]. Currently most trust metrics used in web communities are global, mainly because they are simpler to understand for the users and faster to run on central servers since they have to be executed just once for the entire community. For example Ebay and Pagerank [13] are global. However we think that soon users will start asking for systems that take into account their own peculiar points of view and hence local trust metrics, possibly to be run in a decentralized fashion on their own devices.

While research on trust metrics is quite recent, there have been some proposals for trust metrics. We briefly review some of them for later mention in the evaluation presented in Section 4, although our goal is not to provide a complete review of trust metrics here.

Ebay web site shows the average of the feedbacks received by a certain user in her profile page. This can be considered as a simple global trust metric, which predicts, as trust of A in B, the average of all the trust statements received by B [12].

In more advanced trust metrics trust can be extended beyond direct connections. The original Advogato trust metric [9] is global, and uses network flow to let trust flow from a "seed" of 4 users, who are declared trustworthy a priori, towards the rest of the network. The network flow is first calculated on the network of trust statements whose value is Master (highest value) to find who classifies as Master. Then the Journeyer edges are added to this network and the network flow is calculated again to find users who classify as Journeyer. Finally the users with Apprentice status are found by calculating the flow on all but the Observer edges. The untrusted Observer status is given if no trust flow reached a node. By replacing the 4 seed users for an individual user A, Advogato can also be used as a local trust metrics predicting trust from the point of view of A.

The problem of ranking of web pages in the results of a search engine query can be regarded under a trust perspective. A link from page A to page B can be seen as a trust statement from A to B. This is the intuition behind the algorithm Pagerank [13] powering the search engine Google. Trust is propagated with a mechanism resembling a random walk over the trust network.

Moletrust [12] is a local trust metric. Users are ordered based on their distance from the source user, and only trust edges that go from distance n to distance n+1 are regarded. The trust value of users at distance n only depend on the already calculated trust values at distance n-1. The scores that are lower than a specific threshold value are discarded, and the trust score is the average of the incoming trust statements weighted over the trust scores of the nodes at distance n-1. It is possible to control the locality by setting the trust propagation horizon, i.e. the maximum distance to which trust can be propagated.

Golbeck proposed a metric, TidalTrust [11], that is similar to Moletrust. It also works in a breadth first search fashion, but the maximum depth depends on the length of the first path found from the source to the destination. Another local trust metric is Ziegler's AppleSeed [10], based on spreading activation models, a concept from cognitive psychology.

3 Datasets and trust metrics evaluation

Research on trust metrics started a long time ago, but is somehow still in its infancy. The first trust metric could be even ascribed to the philosopher John Locke who in 1680 wrote: "Probability then being to supply the defect of our knowledge, the grounds of it are these two following: First, the conformity of anything with our own knowledge, observation and experience. Secondly, The testimony of others, vouching their observation and experience. In the testimony of others is to be considered: (1) The number. (2) The integrity. (3) The skill of the witnesses. (4) The design of the author, where it is a testimony out of a book cited. (5) The consistency of the parts and circumstances of the relation. (6) Contrary testimonies" [14]. This quotation can give an idea of how many different models for representing and exploiting trust have been suggested over the centuries. However of course John Locke in 1680 didn't have the technical means for empirically evaluating his "trust metric". Even collecting the required data about social relationships and opinions was very hard in old times. The first contributions in analysis real social networks can be tracked down to the foundational work of Jacob Moreno [15] (see Figure 1) and since then many sociologists, economists and anthropologists have researched on social networks and trust. But the advent of the information age has made it possible to collect, represent, analyze and even build networks way beyond that what is possible with pen and paper. Computer scientists and physicists have become interested in social networks, now that both huge amounts of data have become available and computing power has advanced considerably [4, 5].

At Trustlet.org we have started a wiki to collect information about research on trust and trust metrics. We hope to attract a community of people with interest in trust metrics. We have chosen to use the Creative Commons Attribution license so that work can easily (and legally) be reused elsewhere. Our effort shares the vision of the Science Commons project [1] which tries to remove unnecessary legal and technical barriers to scientific collaboration and innovation and to foster open access to data. We have also started a repository of the software we create for our analysis, written in Python and available as Free Software under the GNU General Public License [2].

We believe the lack of generally available datasets is inhibiting scientific work. It's harder to test a hypothesis if it has been tested on a dataset that is not easily available. The other alternative is testing the hypothesis on synthesized datasets, which are hardly representative of real-world situations. Prior to the proliferation of digital networks data had to be acquired by running face-to-face surveys, which could take years to collect data of a mere couple of hundreds of nodes. The proliferation and popularity of on-line social networks has facilitated acquiring data, and the implementation of standards like XFN and common APIs like OpenSocial opens up new possibilities for research [2]. A more widespread availability and controlled release of datasets would surely benefit research and this is one of the goal behind the creation of Trustlet.

Trust network datasets are are directed, weighted graphs. Nodes are entities such as users, peers, servers, robots, etc. Directed edges are trust relationships, expressing the subjective level of trust an entity expresses in another entity [2].

We think it is important that research on trust metrics follows an empirical approach and it should be based on actual real-world data. Our goal with Trustlet is to collect as many datasets as possible in one single place and release them in standard formats under a reasonable license allowing redistribution and, at least, usage in a research context. At present, as part of our effort with Trustlet, we collected and released datasets derived from Advogato, people.squeakfoundation.org, Robots.net and Epinions.com[3].

We describe in detail the Advogato dataset since our experiments (section 4) are run on it. Advogato.org is an online community site dedicated to free software development, launched in November 1999. It was created by Raph Levien, who also used Advogato as a research testbed for testing his own attack-resistant trust metric, the Advogato trust metric [9]. On Advogato users can certify each other as several levels: Observer, Apprentice, Journeyer or Master. The Advogato trust metric uses this information in order to assign a global certification level to every user. The goal is to be attack-resistant, i.e. to reduce the impact of

[1] Science Commons http://sciencecommons.org

[2] GNU General Public License http://www.gnu.org/licenses/gpl.html

[3] See http://www.trustlet.org/wiki/Trust_network_datasets

attackers [9]. Precise rules for giving out trust statements are specified on the Advogato site. Masters are supposed to be principal authors of an "important" free software project, excellent programmers who work full time on free software, Journeyers contribute significantly, but not necessarily full-time, Apprentices contribute in some way, but are still acquiring the skills needed to make more significant contributions. Observers are users without trust certification, and this is also the default. It is also the level a user certifies another user at to remove a previously expressed trust certification.

For the purpose of this paper we consider these certifications as trust statements. T(A,B) denotes the certification expressed by user A about user B and we map the textual labels Observer, Apprentice, Journeyer and Master in the range [0,1], respectively in the values 0.4, 0.6, 0.8 and 1.0. This choice is arbitrary and considers all the certifications are positive judgments, except for "observer" which is used for expressing less-than-sufficient levels. For example, we model the fact raph certified federico as Journeyer as T(raph, federico)=0.8.

The Advogato social network has a peculiarly interesting characteristic: it is almost the only example of a real-world, directed, weighted, large social network. However, besides the leading work of Levien reported in his unfinished PhD thesis [9], we are just aware of another paper using the Advogato dataset which is focused on providing a trust mechanism for mobile devices [16].

There are other web communities using the same software powering Advogato.org and they have the same trust levels and certifications system: robots.net, persone.softwarelibero.org, people.squeakfoundation.org, kaitiaki.org.nz. We collected daily snapshots of all these datasets and made them available on Trustlet but we haven't used them for our analysis in this paper, mainly because they are much smaller than the Advogato dataset. Details about the characteristics of the Advogato trust network dataset are presented in Section 4.

The other set of datasets we released is derived from Epinions.com, a website where users can leave reviews about products and maintain a list of users they trust and distrust based on the reviews they wrote [12].

Both released datasets and datasets we are considering for collection are available on Trustlet. Besides aiming at releasing datasets in a coherent format, we also released the Python code we wrote for the main trust metrics presented in section 3 and some baseline trust metrics, under a free software license so that code can be reused and inspected.

4 Initial research outcomes

In the previous sections we highlighted the reasons for creating Trustlet and the way we hope it can develop into a collaborative environment for the research of trust metrics. As a first example of what we hope Trustlet will be able to bring to research on trust metrics, we report our first investigation and empirical findings.

We chose to start studying the Advogato social network because of its almost unique characteristic. Trust statements (certifications) are weighted and this makes it a very useful dataset for researching trust metrics: most networks just exhibit a binary relationship (either trust is present or not) and the evaluation on trust metrics performances is less insightful.

The Advogato dataset we analyzed is a directed, weighted graph with 7294 nodes and 52981 trust relations. There are 17489 Master judgments, 21977 for Journeyer, 8817 for Apprentice and 4698 for Observers. The dataset is comprised of 1 large connected component, comprising 70.5% of the nodes, the second largest component contains 7 nodes. The mean in- and out-degree (number of incoming and outgoing edges per user) is 7.26. The mean shortest path length is 3.75. The average cluster coefficient [4] is 0.116. The percentage of trust statements which are reciprocated (when there is a trust statement from A to B, there is also a trust statement from B to A) is 33%.

While a large part of research on social networks focuses on exploring the intrinsic characteristics of the network [4, 5], on Trustlet we are interested in covering an area that received much less attention, analysis of trust metrics. We have compared several trust metrics through leave-one-out, a common technique in machine learning. The process is as follows: one trust edge (e.g. from node A to node B) is taken out of the graph and then the trust metric is used to predict the trust value A should place in B, i.e. the value on the missing edge. We repeat this for all edges to obtain a prediction graph, in which some edges can contain an undefined trust value (where the trust metric could not predict the value). The real and the predicted values are then compared in several ways: the coverage, which is a measure of the edges that were predictable, the fraction of correctly predicted edges, the mean absolute error (MAE) and the root mean squared error (RMSE). Surely there are other ways of evaluating trust metrics: for example it can be argued that an important task for trust metrics is to suggest to a user which other still unknown users are more trustworthy, for example suggesting a user worth following on a social bookmarking site such as del.icio.us or on a music community such as Last.fm (for example because she is trusted by all the users the active user trusts). In this case the evaluation could just concentrate on the top 10 trustworthy users. But in this first work we considered only leave-one-out.

4.1 Evaluation of trust metrics on all trust edges

Table 1 reports our evaluation results of different trust metrics on the Advogato dataset. It is a computation of different evaluation measures on every edge present in the social network. The reported measures are fraction of wrong predictions, Mean Absolute Error, Root Mean Squared Error and coverage. We now describe the compared trust metrics. As already mentioned, we released the code and we plan to implement more trust metrics and release them and run the evaluations. We also applied a threshold function in case of trust metrics that can return values in a continuous interval, such as Moletrust and PageRank, so that for example a predicted trust of 0.746 becomes 0.8 (Apprentice).

The compared trust metrics are some trivial ones used as baselines such as Random, which predicts simply a random trust score in the range [0.4, 1] thresholded in the normal way, or the metrics starting with “Always” which

Table 1. Evaluation of trust metrics on all trust edges

	Fraction wrong predictions	MAE	RMSE	Coverage
Random	0.737	0.223	0.284	1.00
AlwaysMaster	0.670	0.203	0.274	1.00
AlwaysJourneyer	0.585	0.135	0.185	1.00
AlwaysApprentice	0.834	0.233	0.270	1.00
AlwaysObserver	0.911	0.397	0.438	1.00
Ebay	0.350	0.086	0.156	0.98
OutA	0.486	0.106	0.158	0.98
OutB	0.543	0.139	0.205	0.92
Moletrust2	0.366	0.090	0.160	0.80
Moletrust3	0.376	0.091	0.161	0.93
Moletrust4	0.377	0.092	0.161	0.95
PageRank	0.501	0.124	0.191	1.00
AdvogatoLocal	0.550	0.186	0.273	1.00
AdvogatoGlobal	0.595	0.199	0.280	1.00

always return the corresponding value as predicted trust score. Other simple trust metrics are OutA which, in predicting the trust user A could have in user B, simply does the average of the trust statements outgoing from user A, and OutB which averages over the trust statements outgoing from user B.

The other trust metrics were already explained in Section 2, here we just report on how we thresholded and how we run them. Ebay refers to the trust metric that, in predicting the trust user A could have in user B, simply does the average of the trust statements incoming in user B, i.e. the average of what all the users think about user B. MoletrustX refers to Moletrust applied with a trust propagation horizon of value X. The values returned by PageRank as predicted trust follow a powerlaw distribution, there are few large PageRank scores and many tiny ones. So we decided to rescaled the results simply by sorting them and linearly mapping them in the range [0.4, 1], after this we thresholded the predicted trust scores. Our implementation of Advogato is based on Pymmetry1. AdvogatoGlobal refers to the Advogato trust metric run considering as seeds the original founders of Advogato community, namely the users “raph”, “federico”, “miguel” and “alan”. This is the version that is running on the Advogato web site for inferring global certifications for all the users. This version is global because it predicts a trust level for user B which it is the same for every user.

AdvogatoLocal refers to the local version of Advogato trust metric. For example, when predicting the trust user A should place in user B, the trust flow starts from the single seed “user A”. This version is local because it produces personalized trust predictions which depends on the current source user and can be different for different users. AdvogatoLocal was run on a subset (8%) of all the edges since the current implementation is very slow. Due to the leave-one-out technique the network will be different for every evaluation and it has to be restarted from scratch for every single trust edge prediction.

The results of the evaluation are reported in Table 1. We start by commenting the column "fraction of wrong predictions". Our baseline is the trust metric named "Random" which produces an incorrect predicted trust score 74% of the times. The best one is Ebay with an error as small as 35% followed by Moletrust2 (36.57%), Moletrust3 (37.60%) and Moletrust4 (37.71%). Increasing the trust propagation horizon in Moletrust allows to increase the coverage but also increases the error. The reason is that users who are near-by in the trust network (distance 2) are better predictors than users further away in the social network (for example, users at distance 4).

Note that Moletrust is a local trust metric that only uses information located "near" the source node so it can be run on small devices such as mobiles which only need to fetch information from the (few) trust users and possibly the users trusted by them. This behaviour is tunable through setting the trust propagation horizon to specific values. On the other hand, Ebay, being a global trust metric, must aggregate the entire trust network, which can be costly both in term of bandwidth, memory and computation power. The AlwaysX metrics depend on the distributions of certifications and are mainly informative of the data distribution.

The fraction of wrong predictions of Advogato (both local and global) is high compared to Ebay and Moletrust. Advogato was not designed for predicting an accurate trust value, but to increase attack-resistance while accepting as many valid accounts as possible. A side effect is that it limits the amount of granted global certifications and assigns a lot of Observer certificates. In the case of AdvogatoGlobal, 45% of the predicted global certifications are marked as Observer which obviously has an impact on the leave-one-out evaluation. Different trust metrics might have different goals, that require different evaluation techniques. Note that the local version of Advogato is more accurate than the global version. The last metric shown in Table 1 is PageRank [13]: the fraction of correct predictions is not too high but again the real intention of PageRank is to rank web pages and not to predict the correct value of assigned trust.

An alternative evaluation measure is the Mean Absolute Error (MAE). The MAE is computed by averaging the difference in absolute value between the real and the predicted trust statement on an edge. There is no need to threshold values because MAE computes a meaningful value for continuous values. The MAE computed for a certain thresholded trust metric is generally smaller than the MAE computed for the same trust metric when its trust score predictions are not thresholded. But in order to compare metrics that return real values and others that return already thresholded values, we consider the MAE only for thresholded trust metrics. The second column of Table 1 reports the MAE for the evaluated thresholded trust metrics. The baseline is given by the Random trust metric which incurs in a MAE of 0.2230. These results are the worst besides the trivial trust metrics that always predict the most unfrequent certification values. Predicting always Journeyer (0.8) incurs in a small MAE because this value is frequent and central in the distribution. Ebay is the trust metric with the best performance, with a MAE of 0.0855. And it is again followed by Moletrust that in a similar way is more accurate with smaller trust propagation horizons.

A variant of MAE is Root Mean Squared Error (RMSE). RMSE is the root mean of the average of the squared differences. This evaluation measure tends to emphasize large errors, which favor trust metrics that remain within a small band of error and don't have many outlying predictions that might undermine the confidence of the user in the system. For example, it penalizes a prediction as Journeyer when the real trust score should have been Master, or vice versa.

The baseline Random has an RMSE of 0.2839. With this evaluation measure too, Ebay is the best metric with an RMSE of 0.1563 and all the other performances exhibit a pattern similar to the one exposed for the other evaluation measures. However there is one unexpected result: the trivial trust metric OutA is the second best, close to Ebay. Remind that, when asked a prediction for the trust user A should place in user B, OutA simply returns the average of the trust statements going out of A, i.e. the average of how user A judged other users. This trust metric is just a trivial one that was used for comparison purposes. The good performance of OutA in this case is related to the distribution of the data in this particular social setting. The Observer certification has special semantics: it is the default value attributed to a user unless the Advogato trust metric gives a user a higher global certification. So there is little point in certifying other users as Observer. In fact, the FAQ specifies that Observer is "the level to which you would certify someone to remove an existing trust certification". Observer certifications are only when a user changes its mind about another user and wants to downgrade her previously expressed certification as much as possible. This is also our reason for mapping it to 0.4, a less than sufficient level. As a consequence of the special semantics of observer certifications, they are infrequently used. In fact only 638 users used the Observer certification at least once while, for instance, 2938 users used the Master certification at least once. Trust metrics like Ebay and Moletrust work doing averages of the trust edges present in the network (from a global point of view for Ebay and only considering the ones expressed by trusted users for Moletrust) and, since the number of Observer edges is very small compared with the number of Master, Journeyer and Apprentice edges, these predicted average tend to be close to higher values of trust. This means that when predicting an Observer edge (0.4) they lead to a large error. This large error is weighted a lot by the RMSE formula. On the other hand, using the average of the outgoing trust edges (like OutA does) happens to be a successful technique for not incurring in large errors when predicting observer edges. The reason is that a user who used Observer edges tended to use it many times so the average of its outgoing edge certifications is a value that is closer to 0.4 and hence it incurs in lower errors on these critical edges and, as a consequence, in smaller RMSE. This effect can also be clearly seen when different trust metrics are restricted to predict only Observer edges and evaluated only on them. In this case (not shown in Tables), OutA gets the correct value for trust (Observer) 42% of times, while for instance, Ebay only 2.7% of times and

Moletrust2 4%. The fact OutA exhibits a so small RMSE supports the intuition that evaluating which conditions a certain trust metric is more suited for than another one is not a trivial task. Generally knowledge about the domain and the patterns of social interaction is useful, if not required, for a proper selection of a trust metric for a specific application and context.

The last column of Table 1 reports the coverage of the different trust metrics on the Advogato dataset. Sometimes a trust metric might not be able to generate a prediction and the coverage refers to the number of edges that are predictable. The experiment shows that the coverage is always very high. Since local trust metrics use less information (only trust statements of trusted users) their coverage is smaller than the coverage of global trust metrics. Anyway, differently from other social networks [12], it is very high. The Advogato trust network is very dense, so there are many different paths from a user to another user. Even very local trust metrics such as Moletrust2, that only use information from users at distance 2 from the source user, are able to cover and predict almost all the edges.

4.2 Evaluation of trust metrics on controversial users

As a second step in the analysis we concentrated on controversial users [12]. Controversial users are users which are judged in very diverse way by the members of a community. In the context of Advogato, they can be users who received many certifications as Master and many as Apprentice or Observer: the community does not have a single way of perceiving them. The intuition here is that a global average can be very effective when all the users of the community agree that "raph" is a Master, but there can be situations in which something more tailored and user specific is needed. With this in mind we define controversial users as Advogato users with at least 10 incoming edges and standard deviation in received certifications greater than 0.2. Table 2 shows the results of the evaluation of the different trust metrics when they are restricted to predicting the edges going into controversial users. In this way we reduce the number of predicted edges from 52981 to 2030, which is still a significant number of edges to evaluate trust metrics on.

In order to understand better the nature of trust edges under prediction in this second experiment, it is useful to note that, of edges going into controversial users, 1093 are of type Master, 403 of type Journeyer, 115 of type Apprentice and 419 of type Observer. The variance in the values of trust certificates is of course due to the fact that these users are controversial and it is also the reason for which predicting these edges should be more difficult.

We start by commenting the evaluation measures on AlwaysMaster (second row of Table 2) because it presents some peculiarities. Always Master predicts the correct trust value 53.84% (100% 46.16%) of times and, according to the evaluation measure "fraction of correctly predicted trust statements", seems a good trust metric, actually the best one. However the same trust metric, AlwaysMaster, is one of the less precise when RMSE is considered. A similar pattern can be observerd for AdvogatoGlobal. In fact, since in general there

Table 2. Evaluation of trust metrics on trust edges going into controversial users

	Fraction wrong predictions	MAE	RMSE	Coverage
Random	0.799	0.266	0.325	1.00
AlwaysMaster	0.462	0.186	0.302	1.00
AlwaysJourneyer	0.801	0.202	0.238	1.00
AlwaysApprentice	0.943	0.296	0.320	1.00
AlwaysObserver	0.794	0.414	0.477	1.00
Ebay	0.778	0.197	0.240	0.98
OutA	0.614	0.147	0.199	0.98
OutB	0.724	0.215	0.280	0.92
Moletrust2	0.743	0.195	0.243	0.80
Moletrust3	0.746	0.194	0.241	0.93
Moletrust4	0.746	0.195	0.242	0.95
PageRank	0.564	0.186	0.275	1.00
AdvogatoLocal	0.518	0.215	0.324	1.00
AdvogatoGlobal	0.508	0.216	0.326	1.00

is at least one flow of trust with Master certificates going to these controversial users, AdvogatoGlobal tends to predict almost always Master as trust value and since almost half of the edges going into controversial users are of type Master, AdvogatoGlobal often predicts the correct one.

This means that the same trust metric might seem accurate or inaccurate depending on the evaluation measure. This fact once more highlights how evaluating trust metrics on real world datasets is a complicated task and a comparison of same metrics on many different datasets according to different evaluation methods would be highly beneficial for understanding the situation in which one trust metric is more appropriate and useful than another. We already previously explained why OutA is able to have a so small RMSE, the smallest one on controversial users: based on how Observer certifications are used in the system, OutA is the only metric that is able to avoid large errors when predicting the Observer edges.

Arriving at a comparison between a global trust metric such as Ebay and a local trust metric such as Moletrust, we were expecting the latter to be more accurate than the first on controversial users. While on the Epinions dataset, this is what was observed [12], the same is not true here. The reason is partly that in Epinions, the trust values were binary (either trust or distrust) and it was easier to discriminate. Another reason seems to be that on Advogato the user base is not divided in cliques of users such that users of one clique trust each other and distrust users of other cliques. In fact Advogato users are somehow similar and feel part of one single large community. It is future work to analyze if on a social network with a much higher polarization of opinions (such as for example essembly.com, a political site) the performances of local trust metrics are significantly better than global ones.

5 Conclusions

In this paper we have presented Trustlet [7], an open environment for research on trust metrics. We have claimed that the rapid development of social networking asks for a shared effort in collecting datasets and distributing code of algorithms so that comparisons of different research proposals is easier.

As an initial investigation we have reported our comparison of different trust metrics on the Advogato dataset. The results are partly contradictory and this suggests there is need to run systematically evaluations of different algorithms against the same datasets. As future works we are looking into extending our analysis to more datasets also from different social scenarios, for example the networks of relationships (coediting, talk) among Wikipedia users.

Our goal is to make Trustlet an environment which facilitates this collaborative effort. We believe research on these topics is very needed in a time in which our relationships are starting to move more and more into the "virtual" world and our society and life is affected significantly from the predictions and suggestions produced by many different algorithms.

References

1. Francis Fukuyama. Trust: the Social Virtues and the Creation of Prosperity, 1995. Free Press Paperbacks.
2. Paolo Massa. A survey of trust use and modeling in cu rrent real systems, 2006. Chapter in "Trust in E-Services: Technologies, Practices and Challenges", Idea Group, Inc
3. Diego Gambetta, Can We Trust Trust? In "Making and Breaking Cooperative Relations". 2000
4. M. E. J. Newman, The structure and function of complex networks, SIAM Review 45, 167-256 (2003)
5. A.-L. Barabsi Linked: The New Science of Networks (Perseus, Cambridge, MA, 2002)
6. Sabater, J., and Sierra, C., Review on Computational Trust and Reputation Models. Artificial Intelligence Review (2005)
7. Trustlet, collaborative wiki for trust research. http://www.trustlet.org
8. Fullam, K., T. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. Rosenschein, L. Vercouter, and M. Voss. "A Specification of the Agent Reputation and Trust (ART) Testbed: Experimentation and Competition for Trust in Agent Societies" The Fourth International Joint Conference on Autonomous Agents and Multiagent Systems Utrecht, July 2005
9. Raph Levien, Attack Resistant Trust Metrics. Ongoing PhD thesis. http://www.levien.com/thesis/compact.pdf
10. Cai-Nicolas Ziegler. Towards Decentralized Recommender Systems. PhD thesis, Albert-Ludwigs-Universitaet Freiburg, Freiburg i.Br., Germany, 2005

11. Jennifer Golbeck. Computing and Applying Trust in Web-based Social Networks. PhD thesis, University of Maryland, 2005.
12. Paolo Massa, Paolo Avesani, Trust Metrics on Controversial User: Balancing Between Tyranny of the Majority and Echo Chambers. International Journal on Semantic Web and Information Systems 3 (1): 39. (2006)
13. D. Austin, How Google Finds Your Needle in the Web's Haystack, 2006, retrieved on 2008-02-02. http://www.ams.org/featurecolumn/archive/pagerank.html
14. John Locke. An Essay concerning Human Understanding. Harvester Press, Sussex, 1680
15. Jacob Moreno, Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama. Beacon House, Inc. Beacon New York. 1953
16. D. Quercia, S. Hailes, and L. Capra. Lightweight Distributed Trust Propagation. In Proceedings of the 7th IEEE International Conference on Data Mining, 2007