Analyzing User Demographics and User Behavior for Trust Assessment

Davide Ceolin[1], Paul Groth[1], Archana Nottamkandath[1], Wan Fokkink[1], and Willem Robert van Hage[2]

[1] VU University, Amsterdam, The Netherlands
{d.ceolin,p.t.groth,a.nottamkandath,w.j.fokkink}@vu.nl
[2] Synerscope B.V., Eindhoven, The Netherlands
{willem.van.hage@synerscope.com}

Abstract. In many systems, the determination of trust is reduced to reputation estimation. However, reputation is just one way of determining trust. The estimation of trust can be tackled from a variety of other perspectives. In this chapter, we model trust relying on user reputation, user demographics and from provenance. We then explore the effects of combining trust computed through these different methods. Concretely, the first contribution of this chapter is a study of the correlations of demographics with trust. This study helps us to understand which categories of users are better candidates for annotation tasks in the cultural heritage domain. Secondly, we detail a procedure for computing reputation-based trust assessments. The user reputation is modeled in subjective logic based on the user's performance in the system evaluated (*Waisda?* in the case of the work presented here). The third contribution is a procedure for computing trust values based on provenance information, represented using the W3C PROV model. We show how merging the results of these procedures can be beneficial for the reliability of the estimated trust value. We evaluate the proposed procedures and their merger by estimating and verifying the trustworthiness of the tags created within the *Waisda?* video tagging game from the Netherlands Institute for Sound and Vision. Through a quantitative analysis of the results, we demonstrate that using provenance and demographic information is beneficial for the accuracy of trust assessments.

Keywords: Trust, Provenance, Subjective Logic, Machine Learning, Uncertainty Reasoning, Tags

1 Introduction

From deciding the next book to read to selecting the best movie review, we often use the reputation of the author to ascertain the trust in the thing itself. Reputation is an important mechanism in our set of strategies to place trust. In fact, trusting (or placing trust) is an action that we decide or not to perform after having evaluated specific indicators (as specified by O'Hara [25] and also by

Castelfranchi and Falcone, in their theory reprised by Sabater and Sierra [31]), and reputation, i.e. the quantification of a user trustworthiness, is one of these. However, we may base our trust assessment on a variety of other factors as well, including prior performance, a guarantee, or knowledge of how something was produced. Nevertheless, many systems, especially on the Web, choose to reduce trust to reputation estimation and analysis alone. In this work, we take a multi-faceted approach. We look at trust assessment of Web data based on user reputation, provenance (i.e., how data has been produced), and the combination of the two. We also determine the trust on the user based on user profile stereotypes, that are user groups created on the basis of their demographic information. We try to determine correlations between the demographics and the quality of information provided by the users. We use the term "trust" for the trust in information resources and "reputation" for the trust in agents (see the work of Artz and Gil [1] for complete definitions).

We know that over the Web "anyone can say anything about any topic" [35], and this constitutes one of the strengths of the Semantic Web (and of the Web in general), since it brings democracy to it (everybody has the same right to contribute) and does not prevent a priori any possible useful contribution. This makes the Semantic Web a suitable environment for building crowdsourcing platforms. These platforms are useful to collect data (e.g., annotations) from a variety of users, for instance to help cultural heritage and other institutions to classify their collections. However, this brings along trust concerns, since the variety of the contributors can affect both the quality and the trustworthiness of the data. One mechanism for addressing these concerns is to leverage the reputation of the users and the provenance of data.

We perform a series of analyses to demonstrate the existence of correlation between user demographics and identity and the trustworthiness of the data they provide. On the bases of such results, we first propose a procedure for computing reputation that uses basic evidential reasoning principles and is implemented by means of subjective logic opinions [19]. Secondly, we propose a procedure for computing trust assessments based on provenance information represented in the W3C PROV model [34]. Such a procedure is important because it is not always possible to have complete user demographic information. Here, PROV plays a key role, both because of the availability of provenance data over the Web captured using this standard, and because of its role of interchange format: having modeled our procedure on PROV, any other different input format can be easily treated after having been mapped to PROV. We implement this procedure by discretizing the trust values and applying support vector machine classification. Finally, we combine these two procedures in order to maximize the benefit of both. The procedures are evaluated on data provided by the *Waisda?* [24] tagging game[3], where users challenge each other in tagging videos. We show how to use the FOAF ontology to represent the user information provided in their profiles, and we provide a small extension of it to represent user stereotypes. A

[3] A zip file containing the R and Python procedures used, together with the dataset, is retrievable at `http://trustingwebdata.org/books/URSW_III/Waisda.zip`.

stereotype is an abstraction of user demographics. We then provide a procedure to compute the user trustworthiness based on stereotypes from information in user profiles. Through our experiments, we try to determine correlations between the trust of the users and the stereotype of their profile.

We show that a reputation-based prediction is not significantly different from a provenance-based prediction and, by combining the two, we obtain a small but statistically significant improvement in our predictions. We also show that reputation-based and provenance-based assessments correlate and that there is a correlation between the user profile stereotypes and the trust in a user.

This chapter is based on preliminary results published on the paper "Trust Evaluation through User Reputation and Provenance Analysis" [7], presented at the 8th Uncertainty Reasoning for the Semantic Web Workshop at the 11th International Semantic Web Conference 2012. We have revised these results and added an analysis of the correlation between demographics and trustworthiness.

The rest of the chapter is organized as follows: Section 2 describes related work, Section 3 describes the dataset used for evaluation, Section 5, 6, 7 introduce respectively the trust assessment procedures based on reputation, provenance and their combination, including example associated experiments. Section 8 concludes.

2 Related work

Trust is a widely explored topic within a variety of computer science areas. Here, we focus on those works directly touching upon the intersection of trust, provenance, Semantic Web and Web. We refer the reader to the work of Sabater and Sierra [31], Artz and Gil [1], and Golbeck [16] for comprehensive reviews about trust in respectively artificial intelligence, Semantic Web and Web. The first part of our work focuses on reputation estimation and is inspired by the works collected by Masum and Tovey [23]. Pantola et al. [26] present reputation systems that measure the overall reputation of the authors based on the quality of their contribution and the "seriousness" of their ratings; Javanmardi et al. [18] measure reputation based on user edit patterns and statistics. Their approaches are similar to ours, but they are particularly tailored to wiki-based environments. The second part of our work focuses on the usage of provenance information for estimating trust assessments. In their works, Bizer and Cyganiak [2], Hartig and Zhao [17] and Zaihrayeu et al. [38], use provenance and background information expressed as annotated or named graphs [4] to produce trust values. We do not make use of annotated or named graph, but we use provenance graphs as features for classifying the trustworthiness of artifacts. The same difference also applies to the two works of Rajbhandari et al. [30,29], where they quantify the trustworthiness of scientific workflows and they evaluate it by means of probabilistic and fuzzy models. The use of provenance information for computing trust assessments has also been investigated in a previous work of ours [6] where we determined the trustworthiness of event descriptions based on provenance information by applying subjective logic [19] to provenance traces

of event descriptions. In the current chapter, we still represent trust values by means of subjective opinions, but trust assessments are made by means of support vector machines, eventually combined with reputations, again represented by means of subjective opinions. The impact of user information such as age, gender, education and demographics in crowd sourcing tasks have been explored in the works of [20]. In their paper, they explore the relationship between worker characteristics and the quality of their work. Their work has been applied to the crowdsourcing domain and has proven that both the demographics and personality profiles of the workers are strongly linked to the resulting label quality. We apply our algorithm not on a labelling task on a crowdsourcing platform, but on a video annotation task.

Another work by Venanzi et al. [33] addresses the issue of having too few labels from a user to determine their quality by using a community based Bayesian label aggregation model which assumes that crowd workers conform to a few different types, where each type represents a group of workers with similar confusion matrices. We use a similar approach to build stereotypes of users behaviour based on information provided by the users, but not for crowdsourcing systems. Their work is performed on the labeling task while ours is done on annotations of videos. In general, much work has been done in crowdsourcing platforms to determine the effect of a user profile on user accuracy and reputation (see [20], [33]). However, these works focus mainly on labeling crowdsourced data where ground truth data is already available. The main difference between our work on determining correlation of user profiles on their quality with the above mentioned work is that we do not have a ground truth. For the labeling tasks on the crowdsourcing platforms, there is ground truth available for both the works. In our case, we lack such information and thus rely on partial evidence, which is that we trust a tag provided by a user more if there are other users who provided the same tag into the system. Also, the procedure introduced in Section 5 is a generalization of the procedure that we implemented in a few preceding works [8,9,10], where we evaluated the trustworthiness of tags of the Steve.Museum [32] artifact collection.

Lastly, the use of stereotyping as a bootstrapping method has already been investigated by Liu et al. [22] and Burnett et al. [5]. There exist relevant similarities between these works and ours, like, for example, the use of subjective logic to represent trust (this probabilistic logic makes use of Beta and Dirichlet distributions to model trust statistically) and the fact that users can be grouped in stereotypes to obtain useful informations to assess unknown users. Nevertheless, there exist also relevant differences. In fact, both these papers take an agent-approach and their final goal is to determine whether we can trust an agent or not. Our goal, instead, is to determine the agent's (user's) trustworthiness to be able to use it to determine the trustworthiness of the artifact that he or she produces. Also, Burnett et al. propose that agents can learn a stereotyping function, and also Liu et al. propose that stereotyping is based on a function, although they do not investigate it. In our work, we propose to create stereo-

types based on user characteristics (and hence, implicitly, on a function of these characteristics), although we do not explicitly characterize this function.

3 The *Waisda?* dataset

Waisda? [24] is a video tagging gaming platform launched by the Netherlands Institute for Sound and Vision in collaboration with the public Dutch broadcaster KRO. The game's logic is simple: users watch video and tag the content. Whenever two or more players insert the same tag about the same video in the same time frame (10 sec., relative to the video), they are both rewarded. The number of matches for a tag is used as an estimate of its trustworthiness. When a tag which is not matched by others, it is not necessarily considered to be untrustworthy, because, for instance, it can refer to an element of the video unnoticed by other users, or it can belong to a niche vocabulary. Thus, tags that have no matches are not necessarily wrong. In the game, when counting matching tags, typos or synonymity are not taken into consideration.

We validate our procedures by using tag matching to estimate the trustworthiness of tag entries produced within the game. Our total corpus contains 37850 tag entries corresponding to 115 tags randomly chosen. These tag entries correspond to about 9% of the total population. We have checked their representativity with respect to the entire dataset. First, we compared the distribution of each relevant feature that we will use in Section 6 in our sample with the distribution of the same feature in the entire dataset. A 95% confidence level Chi-squared test [28] confirmed that the hour of the day and the day of the week distribute similarly in our sample and in the entire dataset. The typing duration distributions (i.e., distributions of the time employed by users to insert tags) instead, are significantly different according to a 95% confidence level Wilcoxon signed-rank test [37]. However, the mode of the two distributions are the same, and the mean differs only 0.1 seconds which, according to the KLM-GOMS model [3], corresponds, at most, to a keystroke. So we conclude that the used sample is representative for the entire data set. A second analysis showed that, by randomly selecting other sets of 115 tags, the corresponding tag entries are not statistically different from the sample that we used. We used 26495 tag entries (70%) as a training set, and the remaining 11355 (30%) as a test set.

In order to determine the correlation of user profile information with user reputation, we used the data from 17 users who provided information about themselves in their user profiles. The remaining users did not provide their data or chose to remain anonymous. Initially, we tried to cluster the users based on their features such as age, number of contributions etc., and tried to draw conclusions about certain stereotypes. However since we had too few users to draw conclusions based on this approach, we opted, instead, to use standard correlation metrics on our data. We used Pearson correlation for the continuous data such as number of tags provided, number of tags provided which were matched with others and their age. For categorical variables such as gender, we used the point biserial correlation metric.

4 Analysis of correlation between user demographics and data trustworthiness

Demographics is the set of quantifiable statistics about a population. A user profile is a collection of personal information about a given user. In this work, we assume that information collected by aggregation of user profiles represents the demographics of the population.

Here, we try to determine if there is a correlation between the user reputation and demographics in the *Waisda?* system. We use the user reputation as a proxy for data trustworthiness.

Our analysis is performed by grouping users based on their demographics and by identifying a correlation between user groups and the trustworthiness of the artifacts they produced. The drawback of our approach is that the users need to provide their details to the system. Since *Waisda?* is an online game, many users chose to participate as anonymous. We realised that the users who actively returned back to the game are mostly the ones who provided their profile information. This is a good indication of which users will actively participate in the system for a longer time. Another thing to note is that, in general, the users may not provide accurate information about themselves in their profile. However, for the sake of this work, we do not take this possibility into account because the users that provided their personal information in the game are known, and hence their information trusted. Moreover, since we take a statistical approach, information inaccuracies, if any, are compensated. The reason why we investigate the correlation between demographics and data trustworthiness is that we hypothesize that certain categories of users may be better performing than others. For instance, younger users may be more attentive or older users may be more accurate. If that is the case, then the stereotype that we define should help us in identifying groups of users whose performance are higher or lower than others.

4.1 User profiles and their representation

The information in the user profile and other quantitative information derived about a user can help to estimate user reputation. Although different systems gather different types of information from a user, there is an overlap between the most common features such as the age, gender, education, etc. Such information provided by the user can be represented using the FOAF ontology [13]. FOAF provides a representation of the individual user along with his details. Apart from the user provided details, we also derive information such as the number of tags contributed by the user, percentage of tags matched with other users, etc. For representing data that are specific to the tagging environment and system, we do not adopt a standard and we use an ad-hoc representation (the property *ex:contributed_tags* for the number of user contributed tags, and the property *ex:matched_tags* for the number of matched tags for a given user).

In our procedure, we also build groups (or stereotypes) of users who share similar characteristics. In order to form groups of users, we use percentiles for

each characteristic in their profile and derived characteristics. Percentiles help in obtaining an even distribution of the users across different profile characteristics and grouping them in stereotypes. One example of a stereotype can be users who are at least 30 years old and female. In order to represent these groups or stereotypes, we utilize the *group* class of FOAF. The groups are formed based on the information in the individual FOAF profile. Fig. 1 depicts an example of users Alice and Mary who are both females above 30 years of age and belong to the same stereotype. In Fig. 1, the stereotype is represented by an entity of type stereo:stereotype, that is a subclass of the foaf:group class. We propose such a subclass to represent user stereotypes. The fact that we use FOAF and a small extension of it is important, because it eases interoperability with the systems that use this widely adopted ontology.

In the next section, we explain a procedure for predicting the reputation of a user based on the aggregation of the reputations of users within the same stereotype.

4.2 Procedure for analyzing the correlation between user demographics and reputation

In order to evaluate the correlation between user demographics and the trustworthiness of the artifacts that they produce, we developed a procedure that groups users in stereotypes according to their personal information, and we check the existence of correlations between the fact that a given user belongs to a certain stereotype and their reputation.

The procedure is as follows:

$\mathsf{proc}\ \mathsf{reputation\_profile\_prediction}(user, reputation, user\_profile) \equiv$
$attribute\_set := attribute\_selection(user\_profile,)$
$attributes := attribute\_extraction(attribute\_profile)$
$reputation\_levels\_aggregation$
$classified\_testset := classify(testset, trainingset)$

The subprocedures used are described below:

attribute_selection Among all the profile information provided by the user, the first step of our procedure chooses the most significant ones: age and gender. In this process we also distinguish between the categorical variables and the continuous variables. This selection can lead to an optimisation of the computation. As shown in equation 1, the reputation of the user is influenced by the characteristics in his profile.

$$user\_reputation = age \otimes education \otimes gender \otimes salary \otimes \ldots \quad (1)$$

attribute_extraction Apart from the user provided information in the profile, we derive information about the user contributions in the system. This information can be the total number of tags provided, total number of tags

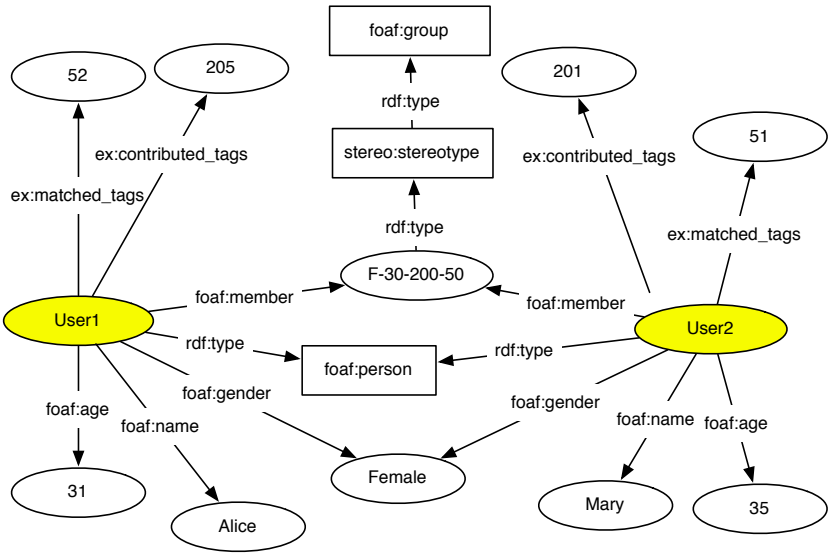

Fig. 1: Graph representation of the users and groups. The group name F-30-200-50 is formed by female users that are older than 30, provided more than 200 tags, and more than 50 of these are matched.

matched with the other users, time spent in the system, etc. This derivation can help us understand the behaviour of the user better and help derive useful correlations about the user behaviour and reputation.

reputation_levels_aggregation To ease the learning process, we aggregate reputation of the users into n classes. The classes are formed by different combinations of features. The features are created based on the extracted user information. To create a feature, we compute percentiles for continuous variables such as age, total tags contributed, etc. Using percentiles, we discretize the continuous variables into four values per feature with each value representing a quarter of the data. However for categorical variables such as gender, education, etc., we use each of the categories available. Once the classes are formed, we consider them as stereotypes of the users. We assign each user to a particular stereotype.

classification Machine learning algorithms (or any other kind of classification algorithm) can be adopted at this stage. The choice can be constrained either from the data or by other limitations (e.g., computational power at our disposal). In this subprocedure, we try to predict information about the reputation of a new user belonging to a certain stereotype based on the reputation of other users belonging to that particular stereotype. This prediction helps to give an "a priori value" of reputation for new users in the system based on information in their profiles.

4.3 Application evaluation

We apply the procedure to the tag entries from the *Waisda?* game as follows.

attribute selection and extraction In the *Waisda?* dataset, we have 17 users who provided in their profile personal information such as e-mail, id, age and gender. The remaining users of the tagging game participated as anonymous. We extract the age and gender from the profiles and derive information such as total number of tags contributed by each user and, for each user, the total number of tags matched with the others. We also compute the reputation of the users using the partial evidence extension of subjective logic that we introduced in a previous work [11] and is summarized as:

$$b = \frac{1}{l+2}\Sigma_{i=1}^{l}\frac{p_i+1}{p_i+2} \quad d = \frac{1}{l+2}\Sigma_{i=1}^{l}\frac{1}{p_i+2} \quad u = \frac{2}{l+2} \tag{2}$$

where b is the belief, d is the disbelief and u is the uncertainty of a subjective opinion. p is a vector of positive observations about distinct facts (e.g., number of matches for different tags provided by the same user). l is the length of p. Each entry in p has a prior probability that is set to the default non-informative value $\frac{1}{2}$. The value of the reputation corresponds to the expected value of the opinion we computed, and is determined as follows.

$$E = b + \frac{1}{2} \cdot u \tag{3}$$

reputation stereotypes computation We discretize the continuous variables such as age, total number of tags contributed, total number of tags matched into four values using percentiles. Each value represents a quarter of the total data. We use this approach to ensure equal distribution of the data for each feature. Categorical variables such as gender represent features that take two values (male, female). Once the features are formed, we aggregate them in different combinations to form the stereotypes of the users. In our case for the *Waisda?* dataset we have 7 stereotypes. We compute a reputation per stereotype based on the evidence at our disposal for the users that belong to it.

regression/classification algorithm We used a regression algorithm to predict the trustworthiness of the users belonging to a stereotype. Once we have sufficient evidence (e.g. at least five or ten users belonging to a stereotype), we can predict the trustworthiness of new users in the system who belong to the same stereotype. This prediction can help us to give an idea about the user trustworthiness in the system and also in the future help to recruit users with certain characteristics for the system.

4.4 Results

Table 1 shows the results of our analysis about the user reputation per stereotype. Here the user reputation is computed by using the formulas presented in

Equation (2) and following Subsection 5.2: for each user in each stereotype we compute the frequency of matched tags that he or she contributed, weighed on the sample size. Here we want to check if stereotypes are able to discriminate users on their reputations. Hence we compute user reputations based on the evidence at our disposal.

Stereotype	# users	User reputations
Stereotype 1	2	[0.96, 0.90]
Stereotype 2	2	[0.97, 0.95]
Stereotype 3	2	[0.91, 0.94]
Stereotype 4	2	[0.97, 0.96]
Stereotype 5	5	[0.97, 0.97, 0.97, 0.98, 0.98]
Stereotype 6	1	[0.94]
Stereotype 7	3	[0.95, 0.93, 0.95]

Table 1: Stereotypes of user profiles and their reputation.

From Table 1, we observe that the maximum variance and maximum standard deviation between the reputations within a stereotype are 0.001 and 0.03 respectively. This shows that there is not much difference between the reputation values of the users belonging to the same stereotype. Also, the difference between stereotypes is very small. However, this small difference may be due to the fact that these stereotypes do not correlate with users reputations. We will investigate in the future the use of stereotypes based only on demographics features that correlate with user reputations to discriminate users on their reputations. Also, in this specific use case, the variance of the user reputation is quite low, so it may be hard to group users based on their reputation. So, instead of checking the correlation between the user stereotype and the user reputation, we evaluate the correlation between user demographics and user reputation, so we decompose the information that determines the user stereotype and we analyze them independently. For data which is normally distributed, we use the Pearson correlation. For categorical data such as age, we use point biserial correlation [15]. The results of our analysis are shown in Table 2.

X	Y	Correlation method	Corr(X,Y)	p-value
# of tags	Reputation	Pearson	0.53	0.02
# of matched tags	Reputation	Pearson	0.61	0.008
Age	Reputation	Pearson	-0.55	0.02
Gender	Reputation	Point biserial	0.46	0.06

Table 2: Results of correlation analysis on *Waisda?* dataset.

From Table 2 we can see that there is linear positive correlation between the number of tags and the number of tags matched with other tags provided by a user with their reputation. However, there is a negative correlation with the user age and their reputation. The point biserial correlation method shows that there is a positive correlation between the gender of the users and their reputation.

Thus, from our experiments, it can be seen that there is a correlation between the information provided by the user and their reputation, at least in the *Waisda?* dataset. For instance, the age correlation indicates that the youngest users perform best, perhaps because they are more reactive and attentive. Also, users that contributed more tags tend to have a higher reputation. This is probably because they developed a better tagging skill over time. Users that contributed a higher number of matched tags also tend to be more precise (it is not given that to a higher number of matched tags corresponds to a higher reputation, since the matched tags could be accompanied by a lot of unmatched ones; this is not the case here). The gender correlation is not significant, since it is even lower than the probability to guess the correct reputation of a user based on his or her gender. These correlations can help us to predict the reputation of new users based on reputations computed from users with similar characteristics. For the moment, these results hold only for this case study, but in the future, we aim to test these features also on additional use cases and to enrich them (both derived from and provided in the profile) to understand how the user characteristics impact the user reputation, and we aim at identifying a corpus of characteristics (shared among use cases) from which we can infer the user reputation. This information may be useful for expert finding, since once we learn which stereotypes of users perform a certain task well, we can recruit more users of that stereotype into the system.

5 Computing reputation-based trust

In the previous section, we analyze some of the assumptions that underpin the use of user reputations for making trust assessments. We find that there exists a moderate correlation between the user demographics and the trustworthiness of the data that the population produces. This leads us to conclude that by virtue of the correlation between user reputation and demographics, demographics can be used as a foundation for trust prediction, although particular countermeasures need to be taken to compensate the fact that the existing correlation is only moderate.

Here, we provide a generic procedure that allows to build a reputation for a user, based on a set of evaluated artifacts (e.g., annotations), and to use it for assessing trust of other artifacts created by him. We build the reputation based on a set of evaluated tags contributed by the user and not on user demographics because we have such evaluations at our disposal, and this allows tailoring the reputation to the specific user. Still, the analysis presented before lays the foundations for the use of user reputation for trust prediction.

5.1 Procedure

We present a generic procedure for computing the reputation of a user with respect to a given artifact produced by him.

```
proc reputation(user, artifact) ≡
  evidence := evidence_selection(user, artifact)
  weighted_evidence := weigh_evidence(user, artifact, evidence)
  reputation := aggregate_evidence(weighted_evidence)
```

Evidence_election Reputation is based on historical evidence, hence the first step is to gather all pieces of evidence regarding a given user and select those relevant for trust computation. Typical constraints include temporal (evidence is only considered within a particular time-frame) or semantics (evidence is only considered when is semantically related to the given artifact). *evidence* is the set of all evidence regarding *user* about *artifact*.

```
proc evidence_selection(user, artifact) ≡
  for i :=1 to length(observations) do
    if observations[i].user = user then evidence.add(observation[i]) fi
```

Evidence_Weighing Given the set of evidence considered, we can decide if and how to weigh its elements, that is, whether to count all the pieces of evidence as equally important, or whether to consider some of them as more relevant. This step might be considered as overlapping with the previous one since they are both about weighing evidence: evidence selection gives a boolean weight, while here a fuzzy or probabilistic weight is given. However, keeping this division produces an efficiency gain, since it allows computation to be performed only on relevant items.

```
proc weigh_evidence(user, artifact, evidence) ≡
  for i := 1 to length(evidence) do
    weighted_evidence.add(weigh(evidence[i], artifact))
```

Aggregate_evidence Once the pieces of evidence have been selected and weighed, these are aggregated to provide a value for the user reputation that can be used for evaluation. We can apply several different aggregation functions, depending on the domain. Typical functions are: *count*, *sum*, *average*. Subjective logic [19], a probabilistic logic that we use in the application of this procedure, aggregates the observations in subjective opinions about artifacts being trustworthy based on the reputation of their authors are represented as follows:

$$\omega(b, d, u) \tag{4}$$

where

$$b = \frac{p}{p+n+2} \quad d = \frac{n}{p+n+2} \quad u = \frac{2}{p+n+2} \tag{5}$$

where b, d and u indicate respectively how much we believe that the artifact is trustworthy, non-trustworthy, and how uncertain our opinion is. p and n

are the amounts of positive and negative evidence respectively. Subjective opinions are equivalent to Beta probability distributions (Fig. 2), which range over the trust levels interval $[0 \dots 1]$ and are shaped by the available evidence.

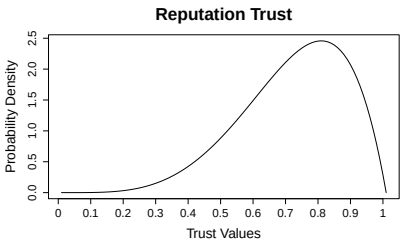

Fig. 2: Example of a Beta probability distribution aggregating 4 positive and 1 negative piece of evidence. The most likely trust value is 0.8 (which is the ratio among the evidence). The variance of the distribution represents the uncertainty about the evaluation.

5.2 Application evaluation

First, we convert the number of matches that each tag entry has into trust values. We obtain an opinion for a given tag entry by aggregating all the evidence (in form of match or non-match) from the other tag entries. For brevity, we report the details about the computation of p and n (i.e. of the positive and negative evidence counts). The corresponding subjective opinion is always computed as in Equation (5).

tag selection For each tag inserted by the user, we select all the matching tags belonging to the same video. In other contexts, the number of matching tags can be substituted by the number of "likes", "retweets", etc..

tag entries weighing For each matching entry, we weigh it on the time distance between the evaluated entry and the matched entry. The weight is determined from an exponential probability distribution, which is a "memoryless" probability distribution used to describe the time between events. If two entries are close in time, we consider it highly likely that they match. If they match but appear in distant temporal moments, then we presume they refer to different elements of the same video. Instead of choosing a threshold, we give a probabilistic weight to the matching entry.
Evidence that $tagentry_i$ contributes to the determination of the trustworthiness of $tagentry$ is represented as $tagentry_{tagentry_i}$. The *timestamp* of $tagentry$ is represented as $t(tagentry)$.

$$p_{tagentry_{tagentry_i}} = exp(y \cdot (t(tagentry) - t(tagentry_i)))$$

$$n_{tagentry_{tagentry_i}} = 1 - p_{tagentry,tagentry_i}$$

where y is a weighing parameter that allows obtaining that 85% of probability mass is assigned to tags inserted in a 10 seconds range (in our case, $y = \frac{1}{5000}$ milliseconds).

tag entries aggregation In this step, we determine the trustworthiness of every tag. We aggregate the weighed evidence in a subjective opinion about the tag trustworthiness. We have at our disposal only positive evidence (the number of matching entries). The more evidence we have at disposal for the same tag entry, the less uncertain our estimate of its trustworthiness will be. Non-matched tag entries have equal probability to be correct or not.

$$p_{tagentry} = \sum_i p_{tagentry,tagentry_i}$$

$$n_{tagentry} = \sum_i n_{tagentry,tagentry_i}$$

We repeat the procedure above for each tag entry created by the user to compute his reputation.

user tag entries selection Select all the tag entries inserted by *user*. We denote a *tagentry* inserted by a *user* as $tagentry_{user}$.

user tag entries weighing Tag entries are weighed by the corresponding trust value previously computed. If an entry is not matched, it is considered as a half positive (tag trust value 0.5) and half negative ($1 - 0.5 = 0.5$) item of evidence (it has 50% probability to be incorrect), as computed by means of subjective opinions. The other entries are also weighed according to their trust value. So, user reputation can either rise or decrease as we collect evidence.

$$p_{tagentry_{user}} = E(\omega_{tagentry})$$

$$n_{tagentry_{user}} = 1 - E(\omega_{tagentry})$$

In the future, we plan to use the reputation the user belongs to as a priori value. In that case, if no items of evidence are available for a user, then his reputation coincides with that of the stereotype he belongs to.

user tag entries aggregation In turn, to compute the reputation of a user with respect to a given tag, we use all the previously computed evidence to build a subjective opinion about the user. This opinion represents the user reputation and can be summarized even more by the corresponding expected value.

$$p_{user} = \sum_{tagentry_{user}} p_{tagentry_{user}}$$

$$n_{user} = \sum_{tagentry_{user}} n_{tagentry_{user}}$$

5.3 Results

We implement the abstract procedure for reputation computation and we evaluate its performance by measuring its ability to make use of the available evidence to compute the best possible trust assessment. Our evaluation does not focus on the ability to predict the exact trust value of the artifact by computing the user reputation, because these two values belong to a continuous space, and they are computed on a different basis. What we expect is that these two values hint at trustworthiness in a similar fashion. We suppose that the trust evaluation system is implemented in such a manner that tags are "accepted" as trustworthy when their trust value is higher than a particular value (also called threshold). So, if the user reputation is a good indicator of trustworthiness, the reputation of a user should be higher than the threshold when the trust values of the artifacts created by him pass the threshold, and vice-versa. The validation, then, depends upon the choice of the threshold which, in turn, depends on constraints imposed by each specific use case. For instance, as we explain below, in the case study we tackle, "false negatives" are preferred over "false positives", and this makes the threshold more likely to be set high (e.g., at least 85% or 90%).

We run the procedure with different thresholds as presented in Fig. 4. Low thresholds correspond to low accuracy in our predictions. However, as the threshold increases, the accuracy of the prediction rises. Moreover, we should consider that: (1) it is preferable to obtain "false negatives" (reject correct tags) rather than "false positives" (accept wrong tags), so high thresholds are more likely to be chosen (e.g., see [14]), in order to reduce risks. Rejecting correct tags means rejecting useful information and therefore wasting part of the effort spent in crowdsourcing tags. Accepting wrong tags means to introduce in the system wrong information and therefore, the tasks that rely on these crowdsourced tags may be affected by this (e.g. if we run an information retrieval task using these tags, then we may retrieve wrong items). Hence we prefer the first situation in place of the latter; (2) a Wilcoxon signed-rank test at 95% confidence level proved that the reputation-based estimates outperform blind guess estimates (having average probability of accuracy 50%). The average improvement is 8%, the maximum is 49%.

We previously adopted this procedure to compute the trustworthiness of tags on the Steve.Museum artifacts [8]. By adapting the procedure to the *Waisda?* case, we were able to formulate the general procedure above.

6 Computing provenance-based trust

User demographics and, in general, user identities are not always available when estimating the trustworthiness of artifacts. Hence, we provide a procedure for estimating the trustworthiness of artifacts based on "how" they were produced rather then on "whom" produced them. Thus, we focus on the "how" part of provenance, i.e., the steps or activities performed in the production of an artifact. (For simplicity, in the rest of the chapter, we will use the word "provenance" to refer to the "how" part). We learn the relationships between PROV and trust

values through machine learning algorithms. This procedure allows to process PROV data and, on the basis of previous trust evaluations, predict the trust level of artifacts.

6.1 Procedure

We present the procedure for computing trust estimates based on provenance.

proc provenance_prediction($artifact\_provenance, artifact$) $\equiv$
  $attribute\_set := attribute\_selection(artifact\_provenance)$
  $attributes := attribute\_extraction(attribute\_set)$
  $trust\_levels\_aggregation$
  $classified\_testset := classify(testset, trainingset)$

attribute_selection Among all the provenance information, the first step of our procedure chooses the most significant ones: agent, processes, temporal annotations and input artifacts can all hint at the trustworthiness of the output artifact. This selection can lead to an optimization of the computation.

attribute_extraction Some attributes need to be manipulated to be used for our classifications, e.g., temporal attributes may be useful for our estimates because one particular date may be particularly prolific for the trustworthiness of artifacts. However, to ease the recognition of patterns within these provenance data, we extract the day of the week or the hour of the day of production, rather than the precise timestamp. In this way we can distinguish, e.g., between day and night hours (when the user might be less reliable). Similarly, we might refer to process types or patterns instead of specific process instances.

trust_level_aggregation To ease the learning process, we aggregate trust levels in n classes. Our results will show that this classification process does not affect accuracy significantly.

classification Machine learning algorithms (or any other kind of classification algorithm) can be adopted at this stage. The choice can be constrained either from the data or by other limitations.

6.2 Application evaluation

We apply the procedure to the tag entries from the *Waisda?* game as follows.

attribute selection and extraction The provenance information available in *Waisda?* is represented in Fig. 3, using the W3C PROV ontology. First, for each tag entry we extract: *typing duration*, *day of the week*, *hour of the day*, *game_id (to which the tag entry belongs)*, *video_id*. This is the "how" provenance information at our disposal. Here we want to determine the trustworthiness of a tag given the modality with which it was produced, rather than the author reputation. Some videos may be easier to annotate than others, or, as we mentioned earlier, user reliability can decrease during the night. For similar reasons we use all the other available features.

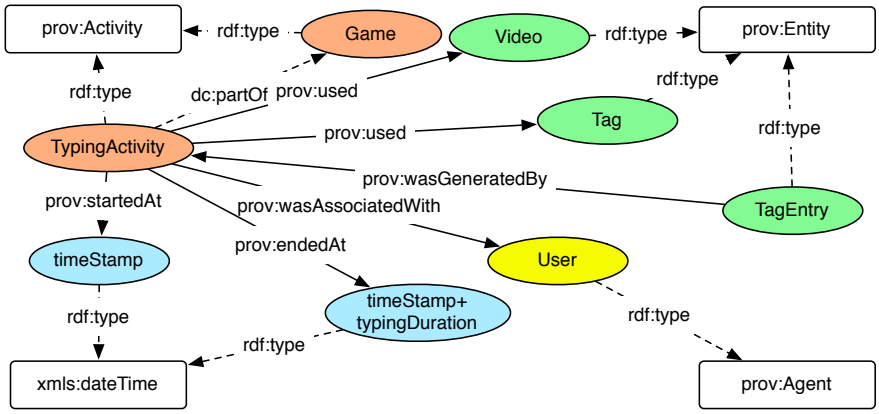

Fig. 3: Graph representation of the provenance information about each tag entry.

trust level classes computation In our procedure, we are not interested in predicting the exact trust value of a tag entry. Rather we want to predict the range of trust values that hold for an entry. Given the range of trust values $[0 \dots 1]$, we split it into 20 classes of length 0.5: from $[0, 0.05]$ to $[0.95, 1.0]$. This allows us to increase the accuracy of our classification algorithm without compromising the accuracy of the predicted value or the computation cost. The values in each class were approximated by the middle value of the class itself. For instance, the class $[0.5 \dots 0.55]$ are approximated as 0.525.

regression/classification algorithm We use a regression algorithm to predict the trustworthiness of the tags. Having at our disposal five different features (in principle, we might have more), and given that we are not interested in predicting the "right" trust value, but the class of trustworthiness, we adopt the "regression-by-discretization" approach [21], that allows us to use the Support Vector Machines algorithm (SVM) [12] to classify our data after having discretized the continuous ones. The training set is composed by 70% of our data, and then we predict the trust level of the test set. We used the SVM version implemented in the e1071 R library [36]. In the future, we will consider alternative learning techniques.

6.3 Results

The accuracy of our predictions depends, again, on the choice of a threshold. If we look at the ability to predict the right (class of) trust values, then the accuracy is about 32% (which still is twice as much as the average result that we would have with a blind guess), but it is more relevant to focus on the ability to predict the trustworthiness of tags within some range, rather than the exact trust value. Depending on the choice of the threshold, the accuracy in this case varies in the range of 40% - 90%, as we can see in Fig. 4. For thresholds higher than 0.85 (the most likely choices), the accuracy is at least 70%. We also compared

the provenance-based estimates with the reputation-based ones, with a 95% confidence level Wilcoxon signed-rank test that proved that the estimates of the two algorithms is not statistically different. *For the* Waisda? *case study, reputation- and provenance-based estimates are equivalent: when reputation is not available or it is not possible to compute it, we can substitute it with provenance-based estimates.* This is particularly important, as the availability of PROV data grows, one can compute trust values for data which is not associated with a trust value.

The "regression-by-discretization" approach consists in first a discretization of the continuous features at our disposal (e.g., timestamps) and a subsequent computation of regression by means of a classification algorithm (e.g., Support Vector Machines). If we apply it for making provenance-based assessments, then we approximate our trust values. This is not necessary with the reputation approach. Had we applied the same approximation to the reputations as well, then provenance-based trust would have performed better, as proven with a 95% confidence level Wilcoxon signed-ranked test, because reputation can rely only on evidence regarding the user, while provenance-based models can rely on larger data sets. Anyway, we have no need to discretize the reputation and, in general, we prefer it because of its lightweight computational overhead.

7 Combining reputation and provenance-based trust

Lastly, we provide a procedure for combining reputation- and provenance-based estimates to improve our predictions. If a certain user has been reliable so far, we can reasonably expect him/her to behave similarly in the near future. So we use reputation and we also constantly update it, to reduce the risk of relying on over-optimistic assumptions (if a user that showed to be reliable once, will maintain his/her status forever). However, reputation has an important limitation. To be reliable, a reputation has to be based on a large amount of evidence, which is not always possible. So, both in case the reputation is uncertain, or in case the user is anonymous, other sources of information should be used in order to correctly predict a trust value. The trust estimate based on provenance information, as described in Section 6, is based on behavioral patterns which have a high probability to be shared among several users. Hence, if a reputation is not reliable enough, we substitute it with the provenance-based prediction.

7.1 Procedure

The algorithm is as follows:

proc provenance_prediction($user, artifact$) $\equiv$
  $q\_ev = evaluate\_user\_evidence(user, artifact)$
  if $q\_ev > min\_evidence$ then $predict\_reputation$ else $predict\_provenance$ fi

evaluate_user_evidence This function quantifies the evidence. Some implementation examples: (1) *count*; (2) compute a subjective opinion and check

if the uncertainty is low enough. As future work we plan to investigate how to automatically determine q_ev and $evaluate\_user\_evidence$.

7.2 Application evaluation

We adopt the predictions obtained with each of the two previous procedures. The results are combined as follows: if the reputation is based on a minimum number of observations, then we use it, otherwise we substitute it with the prediction based on provenance. We run this procedure with different values for both the threshold and the minimum number of observations per reputation. We instantiate the *evaluate_ user_ evidence(user,artifact)* function as a *count* function of the evidence of *user* with respect to a given *tag*.

7.3 Results

The performance of this algorithm depends both on the choice of the threshold for the decision and on the number of pieces of evidence that make a reputation reliable, so we ran the algorithm with several combinations of these two parameters (Fig. 4). The results converge immediately, after having set the minimum number of observations at two. We compared these results with those obtained before. Two Wilcoxon signed-rank tests (at 90% and 95% confidence level with respect to respectively reputation- and provenance-based assessments) showed that *the procedure which combines reputation and provenance evaluations in this case performs better than each of them applied alone*. The improvement is, on average, about 5%. Despite the fact that most of the improvement regards the lower thresholds, which are less likely to be chosen (as we saw in Section 5), even at 0.85 threshold there is a 0.5% improvement. Moreover, we would like to stress how the combination of the two procedures performs better than (in a few cases, equal to) each of them applied alone, regardless of the threshold chosen.

Combining the two procedures allows us to go beyond the limitation of reputation-based approaches. Substituting estimates based on unreliable reputations with provenance-based ones improves our results without significantly increasing risks, since we have previously proven that the two estimates are (on average) equivalent. Hence, when a user is new in a system (and so his/her history is limited) or anonymous, we can refer to the provenance-based estimate to determine the trustworthiness of his/her work, without running a higher risk of poor trust prediction. This improvement is at least partly due to the existing correlation between the reputation and provenance-based trust assessments. A small positive correlation (0.16) has been shown by a Pearson correlation test [27] with a confidence level of 99%. Thanks to this, we can substitute uncertain reputations with the corresponding provenance-based assessments. This explains also the similarity among the results shown in Fig. 4.

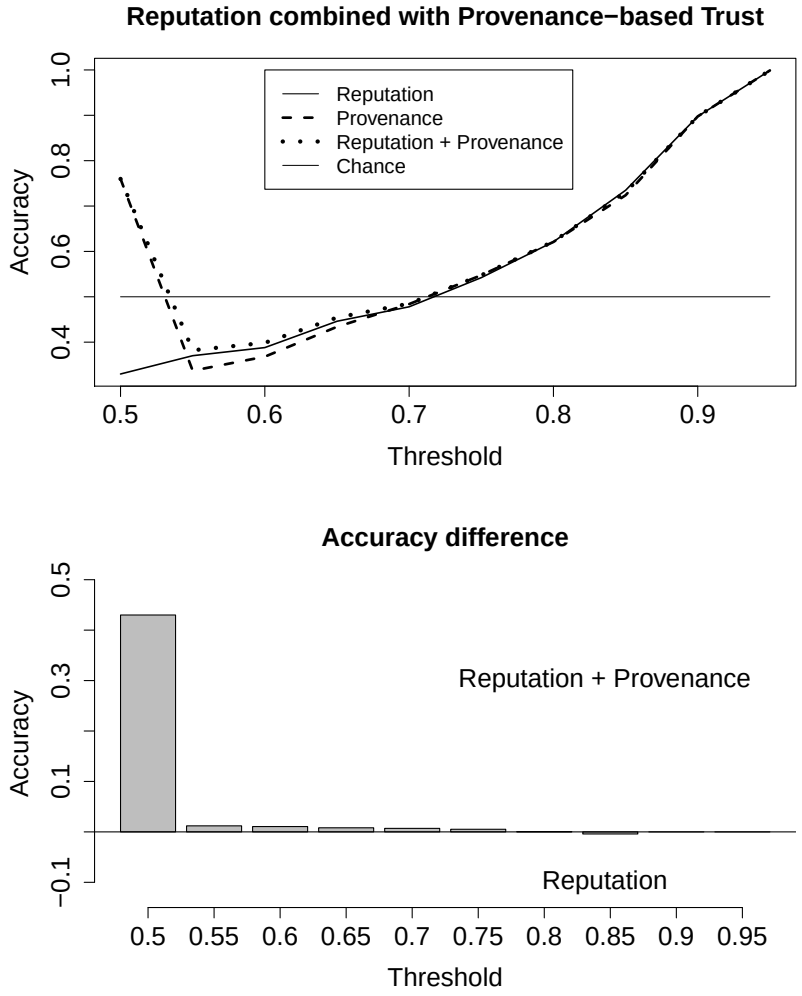

Fig. 4: Absolute and relative (Reputation+Provenance vs. Reputation) accuracy. The gap between the prediction (provenance-based) and the real value of some items explains the shape between 0.5 and 0.55: only very low or high thresholds cover it.

8 Conclusion

In this chapter, we first explored the correlation between user demographics and user reputations and showed the existence of such a correlation in the *Waisda?* tagging dataset. Moreover, we showed how it is possible to use demographics extracted from user profiles to create user stereotypes (user abstractions based on demographics) and to possibly use them as a basis for trust estimation. However, in the *Waisda?* dataset user stereotypes were not useful to discriminate user reputation, although we found a correlation between single demographics (age, gender, etc.) and user reputation. Moreover, we showed how to use the FOAF ontology to both represent user profiles and stereotypes.

Additionally, we proposed and evaluated procedures for computing trust assessments based on reputation, for computing trust assessments based on provenance information, and for combining these two types of assessments. We show that using reputation for trust assessment is simple, computationally light and accurate. We also show the potential of provenance-based trust assessments: these can be at least as accurate as reputation-based methods and can be used to overcome the limitations of a reputation-based approaches (at least within a tagging environment). In *Waisda?* the combination of the two methods was more powerful than each of the two alone. In the future, we will investigate the possibility of automatically extracting provenance patterns usable for trust assessment, to automate, optimize and adapt the process to other case studies and domains. We will also focus on the use of trust assessments as a basis for information retrieval.

Acknowledgements We thank the Netherlands Institute for Sound and Vision for launching and guiding the *Waisda?* project, and our colleagues Michiel, Riste and Valentina for their support. This research was partially supported by the PrestoPRIME project, in the EC ICT FP7 program, and by the Data2Semantics and SEALINC Media projects in the Dutch national program COMMIT.

References

1. D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Journal of Semantic Web*, 5(2):131–197, 2007.
2. C. Bizer and R. Cyganiak. Quality-driven information filtering using the WIQA policy framework. *Journal Web Semantics*, 7(1):1–10, Jan. 2009.
3. S. Card, T. P. Moran, and A. Newell. *The Psychology of Human Computer Interaction*. Lawrence Erlbaum Associates, 1983.
4. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs, provenance and trust. In *WWW '05*, pages 613–622. ACM, 2005.
5. C.Burnett, T.J.Norman, and K.Sycara. Bootstrapping Trust Evaluations Through Stereotypes. In *AAMAS*, pages 241–248. IFAAMAS, 2010.
6. D. Ceolin, P. Groth, and W. R. V. Hage. Calculating the trust of event descriptions using provenance. In *SWPM 2010*, pages 7–12. CEUR-WS, 2010.

7. D. Ceolin, P. Groth, W. R. van Hage, A. Nottamkandath, and W. Fokkink. Trust Evaluation through User Reputation and Provenance Analysis. In *URSW*, pages 15–26. CEUR-WS.org, 2012.
8. D. Ceolin, A. Nottamkandath, and W. Fokkink. Automated evaluation of annotators for museum collections using subjective logic. In *Trust Management VI*, IFIP AICT 374, pages 232–239. Springer, May 2012.
9. D. Ceolin, A. Nottamkandath, and W. Fokkink. Semi-automated assessment of annotation trustworthiness. In *PST 2013*, pages 325–332. IEEE Computer Society, July 2013.
10. D. Ceolin, A. Nottamkandath, and W. Fokkink. Efficient Semi-automated Assessment of Annotation Trustworthiness. *Journal of Trust Management*, 1:1–31, May 2014.
11. D. Ceolin, A. Nottamkandath, and W. J. Fokkink. Subjective logic extensions for the semantic web. In *URSW*, pages 27–38. CEUR-WS.org, 2012.
12. C. Cortes and V. Vapnik. Support-vector networks. *M. Learn.*, 20:273–297, 1995.
13. L. M. Dan Brickley. FOAF. `http://xmlns.com/foaf/spec/`, Jan. 2014.
14. D. Gambetta. *Can We Trust Trust?* Basil Blackwell, 1988.
15. G. V. Glass and K. D. Hopkins. *Statistical Methods in Education and Psychology*. Allyn & Bacon, 1995.
16. J. Golbeck. Trust on the World Wide Web: A Survey. *Foundations and Trends in Web Science*, 1(2):131–197, 2006.
17. O. Hartig and J. Zhao. Using web data provenance for quality assessment. In *SWPM 2009*, pages 26–31. CEUR-WS, 2009.
18. S. Javanmardi, C. Lopes, and P. Baldi. Modeling user reputation in wikis. *Stat. Anal. Data Min.*, 3(2):126–139, Apr. 2010.
19. A. Jøsang. A logic for uncertain probabilities. *Intl. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–212, 2001.
20. G. Kazai, J. Kamps, and N. Milic-Frayling. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *CIKM*, pages 2583–2586. ACM, 2012.
21. I. Kononenko. Naive bayesian classifier and continuous attributes. *Informatica*, 16(1):1–8, 1992.
22. X. Liu, A. Datta, K. Rzadca, and E.-P. Lim. StereoTrust: a Group Based Personalized Trust Model. In *CIKM*, pages 7–16. ACM, 2009.
23. H. Masum and M. Tovey, editors. *The reputation society*. MIT Press, Feb. 2012.
24. Netherlands Inst. for Sound and Vision. Waisda? `http://waisda.nl`, June 2012.
25. K. O'Hara. A General Definition of Trust. Technical report, University of Southampton, 2012.
26. A. V. Pantola, S. Pancho-Festin, and F. Salvador. Rating the raters: a reputation system for wiki-like domains. In *SIN '10*, pages 71–80. ACM, 2010.
27. K. Pearson. Mathematical Contributions to the Theory of Evolution. In *Proceedings of the Royal Society of London*, pages 489–498, 1896.
28. K. Pearson. On the Criterion that a Given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling. *Phil. Mag.*, 50:157–175, 1900.
29. S. Rajbhandari, O. F. Rana, and I. Wootten. A fuzzy model for calculating workflow trust using provenance data. In *MG'08*, pages 1–8. ACM, 2008.
30. S. Rajbhandari, I. Wootten, A. S. Ali, and O. F. Rana. Evaluating Provenance-based Trust for Scientific Workflows. In *CCGRID 06*, pages 365–372. IEEE, 2006.
31. J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24:33–60, 2005.

32. U.S. Institute of Museum and Library Service. Steve Social Tagging Project. http://www.steve.museum/, June 2012.
33. M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *WWW*, pages 155–164, 2014.
34. W3C. PROV-O. http://www.w3.org/TR/prov-o//, June 2012.
35. W3C. Resource description framework (rdf): Concepts and abstract data model. www.w3.org/TR/2002/WD-rdf-concepts-20020829/, June 2012.
36. T. Wien. e1071: Misc functions of the department of statistics (e1071). http://cran.r-project.org/web/packages/e1071/, June 2012.
37. F. Wilcoxon. Individual comparisons by ranking methods. *Biom. Bull.*, 1:80–83, 1945.
38. I. Zaihrayeu, P. da Silva, and D. L. McGuinness. IWTrust: Improving User Trust in Answers from the Web. In *iTrust2005*, pages 384–392. Springer, 2005.