sCARE: Reputation Estimation for Uncertain Web Services

ZAKI MALIK, Wayne State University
BRAHIM MEDJAHED, University of Michigan-Dearborn
ABDELMOUNAAM REZGUI, New Mexico Tech

In this paper, we propose a Statistical Cloud Assisted Reputation Estimation (sCARE) approach for service-oriented environments in uncertain situations. sCARE uses the ratings from cooperating service consumers to uniformly describe the randomness and fuzziness of the different submitted ratings, and their associated relationships in quantitative terms. We also define discriminant functions to model the honesty (or lack thereof) of the service raters. Experiment results show that our proposed model performs in a fairly accurate manner for a number of real-world scenarios. A comparison study with similar existing works is also provided to assess sCARE's performance.

Categories and Subject Descriptors: H.3.5 [**Information Storage and Retrieval**]: Online Information Services - Web services

General Terms: Design, Management, Performance

Additional Key Words and Phrases: Trust, Service Oriented Architecture, Uncertainty

1. INTRODUCTION

Service-Oriented Architecture (SOA) is a development, deployment, and management paradigm that enables an organization to respond to new requirements efficiently. It is a key enabler of the utility computing vision, where consumers subscribe to some provider functionalities (aka services), and pay the providers based on their usage. SOAs thus utilize services as the building blocks for developing software systems distributed within and across organizations. The most common realization of SOAs is based on *Web services*. A Web service is a self-describing software application that can be invoked on the Web using a set of standards (SOAP, REST, etc.) [Noor et al. 2014]. With the introduction of Web services, applications can now be automatically invoked by other Web clients. A primary goal of the Web services technology is therefore enabling the use of Web services as independent components in Web enterprises, that are automatically (i.e., without human intervention) formed as a result of consumer demand and which may dissolve post demand-completion [Medjahed et al. 2003] [Barhamgi et al. 2008] [Benouaret et al. 2014].

Services are defined and designed as self-determining and self-governing components, usually offered by independent service providers. Thus, a services-based enterprise (or a service composition) may involve a number of *a priori* unknown participants, that have little or no prior knowledge of each other, leading to a fairly volatile system in terms of reliability [Papazoglou and Georgakopoulos 2003]. Web services

Author's addresses: Z. Malik, Department of Computer Science, Wayne State University, MI. 48202, zaki@wayne.edu; B. Medjahed, Department of Computer Science, University of Michigan-Dearborn, MI. 48120, brahim@umd.umich.edu; A. Rezgui, Department of Computer Science & Engineering, New Mexico Tech, NM 87801, rezgui@cs.nmt.edu.

DOI: http://dx.doi.org/10.1145/0000000.0000000

may make promises about the provided service and its associated quality but may fail partially or fully to deliver on these promises bringing down the quality of the whole enterprise. In automated services-based transactions, it is therefore imperative that the party that is invoking the other Web service (the consumer) can assess the extent to which the service being invoked (the provider) can provide the required functionality [Malik and Bouguettaya 2009b]. As a plethora service providers are expected to compete in offering similar functionalities, a key requirement is then to accurately assess and estimate the quality of service (QoS) delivered by them [Medjahed and Bouguettaya 2005]. This “assured reliance on the character, ability, or strength” of the service provider is usually referred to as *trust* [Malik and Bouguettaya 2009d]. Establishing trust is therefore a pre-condition for any transaction, and the challenge lies in providing a framework for enabling the selection and composition of Web services based on trust parameters [Noor et al. 2014]. The rationale behind the need for trust is the necessity to interact with unknown entities that have varied QoS delivery levels [Bertino et al. 2004]. There is a growing consensus that the Web service ‘revolution’ would not eventuate until trust related issues are resolved [Birman 2006].

In a service-oriented environment, trust correlates to the ability of a service to perform the required functionality in an acceptable manner [Noor et al. 2013]. The lack of a global monitoring system in SOAs makes trust assessment arduous, and often inaccurate. In addition, the implicitly open and extensive nature of such networked systems means that conventional approaches (e.g., authentication, access control, authorization, etc.) can only provide inadequate and scant solutions for instilling trust in a priori unknown providers. For instance, a provider’s authentication or authorization credentials cannot guarantee that it will exercise these privileges in an expected manner [Malik and Bouguettaya 2009a]. A number of research works show that one can learn from the “wisdom of the crowd” (in the form of ratings and recommendations) for trust estimation [Ben-Naim and Prade 2012] [Noor et al. 2014]. For example, several studies attribute eBay’s commercial success to its reputation mechanism [Houser and Wooders 2005]. Reputation is a subjective assessment of a characteristic or an attribute ascribed to one entity by another based on observations or past experiences. Normally experiences from more than one source (defined as ratings) are assimilated to derive the reputation. This increases the subjectivity of trust and creates *uncertainty*. It has been shown in literature that avoiding unfair ratings to form an unbiased and honest opinion about a service provider is a complex and problematic task [Birman 2006] [Ben-Naim and Prade 2012]. A number of solutions have been proposed in this regard. However, most of these models simply filter, discard or discount the ratings, despite the fact that “reviews are costly and in general not easily obtainable” [Haghpanah and desJardins 2012].

In recent years, theoretical and experimental research has explored the subjective nature of trust. These works are primarily rooted in probability theory, evidence/belief models, or fuzzy logic. Probability based models usually do not consider the element of fuzziness in building trust [Bharadwaj and Al-Shamri 2009] [Shibin et al. 2009]. Since the reasoning is done in a purely statistical manner, they tend to over-formalize trust’s subjectiveness. For example, Bayesian systems take binary ratings as input and assess trust through updating of the beta probability density function [Whitby et al. 2005] [Sabater and Sierra 2003]. This process is fairly complex to comprehend and implement, and loses the component of fuzziness inherent in trust assessment [Noor et al. 2014]. Models based on evidence and belief theory exhibit similar characteristics with added complexity [Jøsang 2001] [Shibin et al. 2009]. On the other hand, fuzzy logic based systems use precise set memberships for defining fuzziness of subjective trust. However, these solutions fail to consider the randomness and uncertainty of membership in those fuzzy sets [He et al. 2004] [Niu et al. 2006]. In this paper, we propose an approach that incorporates the uncertainty and fuzziness in trust estimation to provide a comprehensive and accurate assessment. We use "the statistical cloud model" [Li et al. 2009] in this regard, which integrates these concepts (uncertainty and fuzziness), and their associated relationship in quantitative terms. Specifically,

- We define the 'sCARE model' for reputation assessment, based on three major functions.
- We define an intuitive and mathematically sound function to update rater credibilities.
- We define a reputation function based on the statistical cloud model, to evaluate service provider trust.
- We provide comprehensive experiments and comparison study with similar existing approaches to validate the proposed model.

The paper is organized as follows. In Section 2, we provide a brief overview of the proposed model and related background knowledge. Sections 3, 4, and 5 describe the underlying components of our solution. Section 6 lists the experiments and verifies the applicability of our proposed model in relation to some existing works. Section 7 provides a brief overview of the related work, while Section 8 concludes the paper.

2. THE PROPOSED APPROACH

In this section, we define our proposed approach that consolidates and integrates the uncertain and fuzzy ratings submitted by different service raters to provide a unified and holistic trust assessment of a given service provider. Our solution employs "the statistical cloud model" which defines a way for modeling the transition between a linguistic term of a qualitative concept and its quantitative representation under uncertain and fuzzy conditions.

2.1. Background: The Statistical Cloud Model

The basis of the statistical cloud model (or simply, the cloud model) is that fuzziness and randomness are complementary and essentially inseparable concepts when considered in linguistic terms. It states that the concept of fuzzy membership functions is not sufficient for representing the uncertainty and imprecision in real world settings, and probability theory needs to be incorporated to overcome this inadequacy. In essence, a cloud model can uniformly describe the concepts of randomness, fuzziness, and their relationship in quantitative terms. Experiment results have shown that the cloud model exhibits higher levels of simplicity and robustness in comparison with traditional fuzzy logic and probability based methods [Li et al. 1998]. In the following, we provide a brief overview of the cloud model. For further details, the interested reader is referred to [Li et al. 2009].

Let U be the quantitative universe of discourse, and C denote a qualitative concept associated with U. If $a \in U$ is a random realization of C, and $\mu(a) \in [0,1]$ is a random variable with stable tendency denoting the degree of certainty for a belonging to C:

$$\mu \colon U[0,1] \qquad \forall a \in U \qquad a \to \mu(a)$$

The distribution of a in U is called the cloud (denoted $C(A)$) and each a is called a cloud drop. Note that in probabilistic terms, $a \in U$ is not a simple random number but it has a certainty degree, which itself is also random and not a fixed number. The cloud is composed of a number of drops, which are not necessarily ordered. The underlying character of the qualitative concept is expressed through all cloud drops. Hence the overall feature of the concept is more precisely represented by a large number of drops. The certainty degree of each cloud drop defines the extent to which the drop can represent the concept accurately. Formally, a cloud's quantitative representation is defined over a set of N ordered pairs (a_i, b_i), where a_i is a cloud drop, and b_i is its certainty degree, with $1 \leq i \leq N$.

A one-dimension normal cloud model's qualitative representation can be represented by a triple of quantitative characteristics: Expected value (Ex), Entropy (En) and Hyper-Entropy (He). Ex is the expectation of the cloud drops' distribution, i.e., it corresponds to the center of gravity of the cloud (containing elements fully compatible with the qualitative concept). En represents the uncertainty measurement of a qualitative concept. It is determined by both the randomness and fuzziness of the concept. En indicates how many elements could be accepted to the qualitative linguistic concept. He is a measure of the dispersion on the cloud drops. It can also be considered as En's uncertainty. Vector $\vec{v} = (Ex, En, He)$ is called the eigenvector of a cloud [Li et al. 1998]. The transformation of a qualitative concept expressed by Ex, En, and He to a quantitative representation expressed by the set of numerical cloud drops is performed by the forward cloud generator [Li et al. 2009]. Given these three digital characteristics (Ex, En, He), and the number of cloud drops to be generated (N), the forward cloud generator can create these N cloud drops in the data space with a certainty degree for each drop that each drop can represent the qualitative concept. The procedure is: (1) Generate a normally distributed random number F with mean En and standard deviation He. (2) Generate a normally distributed random number a with mean Ex and standard deviation F. (3) Calculate $b = e^{-\frac{(a-Ex)^2}{2(F)^2}}$. (4) (x, y) represents a cloud drop in the universe of discourse. (5) Repeat Steps 1-4 until N cloud drops are generated. Using this algorithm, the quantitative value of the cloud drops is thus determined by the standard normal form distribution function. Hence, the certainty degree function adopts a bell-shaped curve. This is similar to the one adopted in fuzzy set theory. As mentioned earlier, the normal cloud model is therefore an inclusive model based on probability theory and fuzzy set theory, and is able to depict randomness in the former and fuzziness in the latter. For these reasons we build our trust estimation solution upon the statistical cloud model. For further details on the model and accompanying examples, the interested reader is referred to [Li et al. 2009].

2.2. The sCARE Model: An Overview

We propose a statistical cloud assisted reputation estimation (sCARE) model to estimate provider trust. sCARE is distributed in nature, where in contrast to third-party-based traditional approaches for trust management, no single entity is responsible for collecting, updating, and disseminating ratings provided by different consumers. Each service consumer records its own perceptions of the reputation of only the services it actually invokes. For each service s_i that it has invoked, a service consumer j maintains a rating R_{ij} representing j's perception of s_i's behavior. Different strategies may be adopted in updating R_{ij} which represents only consumer j's perception of the provider s_i's reputation. Other service consumers may differ or concur with j's observation of s_i. A service consumer that inquires about the reputation of a given service provider from its peers may get various R_{ij} "feedbacks." To estimate s_i's behavior, all these feedbacks need to be aggregated. Assume L denotes the set of service consumers which have interacted with s_i in the past and are willing to share their ratings. We assume that L is not empty, i.e., some service willing to share information can be found:

$$Trust(s_i) = \bigwedge_{x \in L} (R_{ix}) \tag{1}$$

where $\bigwedge$ represents the aggregation function. Equation 1 provides a first approximation of how the trust may be assessed. However, it involves a number of factors that are discussed in the following.

Let $R \in \mathbb{R}^n$ be the ratings matrix, $v \in \mathbb{R}^n$ be the majority ratings vector, and $w \in \mathbb{R}^m$ be the credibility vector of the respective raters. We define a *credibility* value for each rater to indicate how much weightage should be given to the rater's reported rating (in terms of the service provider's trust estimation). The entry R_{ij} then represents the rating for a service provider s_i assigned/submitted by the rater j: $R = [r_1 ... r_n]$, where r_i lies in the interval [0,1]. The service providers, raters and their corresponding ratings form a bipartite graph which can be represented using an $n \times m$ adjacency matrix A; i.e., $A_{ij} = 1$ if s_i has a rating from rater j, and 0 otherwise. For conciseness here we assume that each s_i has received a rating from every rater, i.e., $A_{ij} = 1 \; \forall i, j$. Note that in reality this will hardly be the case, and the A_{ij} will be a sparse matrix at best. The difference (d) between a rater j's submitted rating is defined as the Euclidean distance to the majority rating vector v:

$$d = |r_j - v| \tag{2}$$

Equation 2 provides an estimate of each rater's variance compared to the majority rating vector. The proposed model is then governed by three primary functions:

(1) the majority function J : $\mathbb{R}^m \rightarrow \mathbb{R}^m : J(r) = v$,
gives the majority rating vector depending on the submitted ratings R;

(2) the credibility function G : $\mathbb{R}^n \rightarrow \mathbb{R}^m_{\geq 0} : G(v) = w$,
gives the credibility weight vector for the raters depending on d as defined in Eq. 2.

(3) the reputation function F : $\mathbb{R}^m \rightarrow \mathbb{R}^n$: F (w) = $\Re$,
gives the service provider reputation, based on the credibility weights of the raters, and the rating matrix R.

3. THE MAJORITY FUNCTION IN SCARE

A number of successful online shopping stores (e.g., Amazon, eBay, YAHOO! Shopping, etc.) use customer feedbacks to indicate provider ability / dependability. These reputation systems are centralized in nature that mainly rely on the numerical values obtained from the diverse customers. The numerical values are then aggregated to obtain a single 'trust' value, which may not precisely conclude the trustworthiness of the providers. In some systems, the numerical trust value is also complemented with some textual explanation. For instance, in eBay (a standout reputation system) the purchasers and vendors can rate one another on a three point scale, with +1 for a positive rating, 0 for impartial, and -1 for a negative rating. The transaction participants are also asked to leave a textual feedback rating. However, the eBay reputation framework computes trust as a summation of all negative and positive ratings received. Undoubtedly, such evaluations are not always precise. Consider a seller with 50 positive feedbacks, and another with 300 positive and 250 negative ratings. Using eBay's model, both will end up with the same trust rating (+50) [Malaga 2001]. Since humans are involved directly in processing the provided information (reputation value plus textual feedback), this model has prospered [Resnick and Zeckhauser 2002] [Houser and Wooders 2005]. The failure of automated frameworks to reason in a human-like manner implies that such literal feedbacks may not prove completely fruitful, and an eBay-like framework may not be pragmatic for services-based interactions. Similarly, some other online businesses (e.g., Amazon) use a simple-averaging model. Although this method is an improvement, it still does not accurately reflect the reputation as seen in the real world. For instance, a combination of very high and very low ratings would yield the same average as one with all moderate ratings. Therefore, it is well-accepted that defining an evaluations framework that is sufficiently powerful to recognize and alleviate the impacts of variable ratings is critical [Whitby et al. 2005].

In regards to the aforementioned issues, a number of research works screen the ratings based on their divergence from the majority opinion. Some notable mentions include Beta Deviation Feedback [Buchegger and Boudec 2004], Beta Filtering Feedback [Whitby et al. 2005], Likemindedness [Walsh and Sirer 2005], and Entropy-Based Screening [Weng et al. 2005]. In this work, we use the same principle to mitigate discrepancies originating due to unfair or conflicting rater testimonies. The underlying proposition is to increase a rater's credibility (and hence weight of the reported rating) if the reported rating agrees with the majority of the ratings that are reported for that particular provider in a specific time instance, and decreased otherwise. Unlike majority of existing frameworks, we assume that divergence from the majority may occur due to an original experience. Thus, we do not simply disregard or discard the rating, and only modify the credibility value, but still incorporate the diverging opinion. We use a data clustering technique to define the majority opinion by grouping similar feedback ratings together[Vu et al. 2005]. The "majority refinement" (MR) property in our model then states that *ratings that differ from the majority opinion should be assigned less weight*. We use the k-mean clustering algorithm on all current reported ratings to create the clusters. The most densely populated cluster is then labeled as the "majority cluster" and the centroid of the majority cluster is taken as the *majority rating* (denoted v):

$$v = centroid(max(\Gamma_k)) \qquad \forall k \tag{3}$$

where k is the total number of clusters, $max()$ gives the cluster Γ with the largest membership and $centroid()$ gives the centroid of a given cluster.

4. THE CREDIBILITY FUNCTION IN SCARE

The foremost drawback of feedback-only based systems is that all ratings are assumed to be honest and unbiased. For instance, in an averaging model, the feedbacks from all raters are weighed equally so the final service provider reputation comes out as a simple average of the submitted ratings. Similarly, in PageRank-based techniques where a random walk of the network is accepted as a viable model of Web surfing, trust assessment equates to the acceptance of trust transitivity, i.e., if X trusts Y and Y trusts Z, then it is implied that X should trust Z. However, in the real world we clearly distinguish between the testimonies of our sources and weigh the "trusted" ones more than others [Tennenholtz 2004]. A Web service that provides satisfactory service (in accordance with its promised quality (QoS_p)), may get incorrect or false ratings from different evaluators due to several malicious motives. In order to cater for such "bad-mouthing" or collusion possibilities, a trust framework should weigh the ratings of highly credible raters more than consumers with low credibilities [Delgado and Ishii 1999] [Xiong and Liu 2004] [Malik and Bouguettaya 2009b].

In our model, the final trust value is calculated according to the credibility scores of the raters (used as the weight). The credibility weight for the rater is updated at each reputation assessment instance, and is weighted according to the existing rater credibility. Hence, we can write function G as:

$$G(v) = \begin{bmatrix} w_1 \pm g(d_1) \\ \vdots \\ w_m \pm g(d_m) \end{bmatrix} \tag{4}$$

where w_j is the credibility of rater j, $\pm$ indicates that w is either increased or decreased by the factor obtained from the discriminant function g. G satisfies the MR-property if its associated discriminant function $g : \mathbb{R} \rightarrow \mathbb{R}$ is non-negative. Moreover, for the case where the credibility weight is to be increased (+), G would be monotonically increasing, and decreasing otherwise. In [Laureti et al. 2006], Laureti et al. define a discriminant function $g(d) = d^k$ (with $k \geq 0$). We propose $g(d) = w^t \times \theta \times (1 - d_j)$, and $g(d) = e^{-\tau d}$ as the two discriminant functions (to be discussed later) for the credibility weight increase and decrease respectively. Since all transactions may not be equally weighed in terms of their importance, a service consumer may decide to decrease a dishonest rater's credibility according to the transaction's "impact". The transaction impact factor (τ) lies in the range [0, 1] and is assigned a high value for high impact transactions, and vice versa by the service consumer.

DEFINITION 1: The quadratic MR system is a system of equations in the reputations r^t of service providers, and the credibility weights w^t of raters that evolve over discrete time t according to the rating matrix R, with starting weights potentially being equal:

$$r^t = F(w^t) = \frac{R \times w^t}{\sum_{j=1}^{m} w_j} \tag{5}$$

$$w^t = G(v^t) \tag{6}$$

Note that in Def. 1 above, we have defined $F(w)$ as simply a weighted average of all the ratings. This is only done for the sake of simplicity, and in this paper we use the statistical cloud model (defined previously) instead.

DEFINITION 2: The proposed model is a quadratic MR system, with a positive transaction impact τ, and the discriminant function(s) $g(d)$ with

$$w^{t+1} = \begin{cases} w^t + \{w^t \times \theta \times (1 - d_j)\} & d_j < \theta \\ w^t \times e^{-\tau d_j} & \text{otherwise} \end{cases} \tag{7}$$

where w lies in the interval [0, 1] with 0 identifying a completely dishonest rater and 1 an honest one. θ is a pre-defined threshold for acceptable d values. If d is less than θ, it means that the rater's submitted rating is 'close' to the majority opinion. Per Definition 2 above, the rater credibility is thus increased in a linear manner. Otherwise, the rater's credibility is decreased exponentially by a factor of d, i.e., greater the d, more the rater credibility will decrease. This is in accordance with the sociological trust building process where it is difficult to gain inter-personal trust, but easy to lose it.

Let us now look at the two discriminant functions defined above. We can see that Eq. 7 forms a quadratic MR system, which can lead to another similar system where τ is not the transaction importance factor, but a number based on the variance of ratings. The probability density function f for ratings r_j with a given v and the variance in the ratings σ is given then by:

$$f(r_j | v, \sigma) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(R_{ij} - v)^2}{\sigma^2}}$$

Note that the above applies when $d_j \geq \theta$, i.e., the distance from the majority rating in the current time instance is above the acceptable threshold. For simplicity, we assume that σ holds the same value for all raters, independent of their d_j. If we do not base the credibility values of raters on their past behavior (as done in Equation 7), then the same variance may lead to limitations in the model. However, since w^{t+1} depends on w^t in our model, this is an acceptable assumption, and a more complex method is not strictly desired. We can recover the quadratic MR system with $g(d) = e^{-\tau d}$ by considering the iterations on the ratings, and those on the weights as:

$$w^{t+1} = \begin{bmatrix} f(r_1|v,\sigma) \\ \vdots \\ f(r_m|v,\sigma) \end{bmatrix} = (\sqrt{2\pi}\sigma)^{-n} \begin{bmatrix} e^{-\frac{n}{\sigma^2} d_1^{t+1}} \\ \vdots \\ e^{-\frac{n}{\sigma^2} d_m^{t+1}} \end{bmatrix}$$

From above we can see that τ in this case turns out to be proportional to the inverse of the variance, i.e., $\frac{n}{\sigma^2}$. This implies that if σ is large enough, τ is decreased, and varying rater opinions are accepted more, leading to g being less discriminating and hence w^{t+1} decreasing at a slightly lower rate (but still monotonically decreasing).

The discriminant function from Equation 7 (for $d_j < \theta$) is based on an affine function involving the distances and past credibilities of raters, and the associate system has a statistical interpretation. For this purpose, we can use the log-likelihood of the probability density function above. This is similar to the degree of belief for each rater used in [Zhang et al. 2006]. The MR system with $g(d) = w^t \times \theta \times (1 - d_j)$ may then be obtained by considering:

$$w^{t+1} = log \begin{bmatrix} f(r_1|v,\sigma) \\ \vdots \\ f(r_m|v,\sigma) \end{bmatrix}$$

where log is to be applied component-wise. The objective function that the system maximizes, can then be defined as $\sum_j w^t d_j - \frac{\theta}{2} \sum_j w^t d_j^2$. The first term is merely a weighted average of the submitted ratings and hence maximized as such, while the latter is maximized by taking v in $[0,1]^n$. Therefore, the proposed solution provides a compromise between the average and the solution that diverges from this average (using the majority rating). We can also analyze the definition of r, as given in Equation 5- 7 and its convergence property. It can be seen that the scalar function $\zeta : [0,1]^m \to \mathbb{R} : r \mapsto \zeta(r)$ defined in these equations is continuous. If r^t and r^{t+1} are two successive iterations of r, then we can state that the two terms are associated with each other as

$$r^{t+1} = r^t + \alpha.grad\zeta(r^t)$$

where $\alpha > 0$ is a constant, and $grad$ denotes the gradient of ζ in r^t pointing to the greatest ascent direction (respectively descent, for decreasing credibilities). Thus, one iteration corresponds to taking the direction of greatest ascent/descent with a step length $l^t = \alpha.||grad\zeta(r^t)||_2$. Moreover, the strictness of the ascent (corresponding descent) can be shown for l^t, such that $\zeta(r[t]) < \zeta(r[t+1])$ (and $>$ vice versa). Let c be a constant, such that $\alpha \geq c > 0$, then l^t is also lower (alt. upper) bounded by $c||grad\zeta(r[t])||_2$, and therefore monotonically converges on r to the maximizer (alt. minimizer) of ζ on $[0,1]^m$.

Credibility Bootstrapping: In cases where no Cr^t exists, i.e., the rater and consumer have not previously interacted, the rater's initial credibility is set at the middle (0.5) to indicate impartiality (on a scale from 0 to 1). However, previous research has shown that assigning pre-defined high or average values may encourage "reputation white-washing". Interested readers are referred to [Malik and Bouguettaya 2009c] for a detailed discussion, and alternative models for bootstrapping rater credibility. Therefore, we dampen the bootstrap value by weighing in the consumer's pessimistic/optimistic preferences towards services interactions, i.e., $Cr_{bootstrap} = 0.5 \times \lambda$;

where λ denotes the consumer's pessimistic/optimistic preference in the range [0, 1]. A high λ value indicates an optimistic consumer, one that is willing to trust the testimony of a new rater. Alternatively, $\lambda \leq 0.5$ indicates a pessimistic consumer. The choice of λ is at the discretion of the service consumer. However, to provide a better estimate of the consumer's propensity to accept, we set λ as the *ratio* of the total number of times the ratings submissions (by all raters) are deemed useful (k) by the service consumer, over the total number of rating submissions received by the service consumer (n) [Lam and Riedl 2004]. The λ factor is: $\lambda = \frac{\sum_{i=1}^{k} U_i}{\sum_{x=1}^{m} R_x}$ where U_i is the submission where the rater was termed honest (i.e., $d \leq \theta$) and R_x denotes the total number of rating submissions.

Credibility Update: Note that $g(d)$ as defined in Equation 7 provides w to be used in the reputation function (defined below). However, its value is not final, i.e., w is also updated after the service requester's personal experience ($p_1 ... p_m$) for that time instance. If the difference between p_j and each rater i's submitted rating R_{ij} falls under a pre-determined threshold β, it is left as is, otherwise it is decremented by $\gamma \in [0, 1]$. The choice of γ is left at the discretion of the service requester. The lower the value, the more pessimistic is the consumer and higher values of γ are suitable for optimistic consumers. We define a pessimistic consumer as one that does not trust the raters easily and reduces their credibility drastically on each false feedback. On the other hand, optimistic consumers tend to forgive dishonest feedbacks assuming that the difference in opinion could be due to environmental factors (e.g., network delay, etc.), or a "one-off" instance. Then:

$$w^{t+1} = \begin{cases} w^{t+1} & |p_j - R_{ij}| < \beta \\ w^{t+1} \times \gamma & \text{otherwise} \end{cases} \tag{8}$$

5. THE REPUTATION FUNCTION IN SCARE

As mentioned earlier, the credibilities of the raters (calculated in the previous step) are used to assess the service provider's reputation. Since service provider ratings decay with time all the past reputation data may be of little or no importance [Malik and Bouguettaya 2009b], [Marti and Garcia-Molina 2004]. For instance, a Web service performing inconsistently in the past may ameliorate its behavior. Alternatively, a service's performance may degrade over time. It may be the case that considering all historical data may provide incorrect reputation scores. In order to counter such discrepancies, we incorporate temporal sensitivity in our proposed model. The rating submissions are time-stamped to assign more weight to recent observations and less to older ones. This is termed as "reputation fading" where older perceptions gradually *fade* and fresh ones take their place. We adjust the value of the ratings as:

$$R_{ij}^{t+1} = R_{ij}^{t} * f_d \tag{9}$$

where R_{ij} is as defined above and f_d is the reputation fader. In our model, the recent most rating has the fader value 1 while older observations are decremented for each time interval passed. When $f_d = 0$, the consumer's rating is not considered as it is outdated. The "time interval" is an assigned factor, which could be anywhere from a single reputation inquiry, ten inquiries or even more than that. All inquiries that are grouped in one time interval are assigned the same fader value. In this way, the service consumer can define its own temporal sensitivity degree. For example, a service can omit the fader value's effect altogether by assigning it a null value. We propose to use a fader value that can then be calculated as: $f_d = \frac{1}{\sqrt{P_u}}$, where P_u is the time interval difference between the present time and the time in which the rating was collected from the rater. This allows the convergence of reputation to a very small value as time passes. Note that the consumer can assign a group of ratings collected at different times to have the same time-stamp, and hence lie in the same time interval. Other complex values for the fader are also acceptable.

Characteristics Extraction: In the statistical cloud model, the backward cloud generator allows transformation of the cloud model from its quantitative representation to a qualitative one. We incorporate rater credibility values and majority rating to produce the three digital characteristics of the cloud (Ex, En, He). Given a set of N ratings $R_{ij}(j = 1, 2, ..., N)$, we can extract the three characteristics as:

(1) Update R_{ij} values using f_d, for all ratings (including previous time instances).
(2) For each rater j, update w_j (using equations defined previously).
(3) Calculate

$$Ex = \frac{\sum_{j=1}^{N}(w_j R_{ij})}{\sum_{j=1}^{N} w_j}$$

(4) Calculate

$$En = \sqrt{\frac{\pi}{2}} \times \frac{\sum_{j=1}^{N} w_j |R_{ij} - Ex|}{\sum_{j=1}^{N} w_j}$$

(5) Calculate

$$He = \sqrt{\frac{\sum_{j=1}^{N} w_j (R_{ij} - Ex)^2}{\frac{(N'-1)\sum_{j=1}^{N} w_j}{N'}} - (En)^2}$$

where N' is the number of non-zero credibilities.

Reputation Estimation: The next step is using the three discovered characteristics to make a subjective assessment of the provider's trust. Since He is a measure of En's uncertainty, we only use Ex and He to quantify the provider's trust and the associated uncertainty. This allows us to consider the latest majority view of the provider's reputation and the decentralization of ratings from it. A higher value of Ex therefore indicates high reputation, while a small He indicates the stability of the ratings around this decision. Intuitively this makes sense, but for a large N, making these comparisons is non-trivial. For instance, Ex and He can occur together in one of four forms: one is high/low the other is low/high, both are high, or both are low. Therefore, to quantify the relationship between the two characteristics, i.e., the provider (s_i)'s trust assessment, we use:

$$Trust(s_i) = \begin{cases} 1 - \frac{He}{Ex+He} & if Ex \neq 0 \ \& \ He > \theta; \\ Ex & if He \leq \theta; \\ 0 & if Ex = 0; \end{cases}$$

sCARE Example: Here, we discuss a hypothetical scenario for reputation estimation using the proposed sCARE approach. Assume that a consumer *x* wants to conduct business with a service provider s_i. x queries a number of services for s_i's reputation and receives a ratings vector of {0.5, 0.2, 0.5, 0.2, 0.5, 0.5, 0.6, 0.7, 0.3, 0.4} from raters {j=1 through 10}. Moreover, assume that all these raters have prior credibilities of {0.6, 0.6, 0.5, 0.1, 0.3, 0.5, 0.4, 0.5, 0.35, 0.35} stored with x. In a simple weighted-average scheme, s_i's reputation will be 0.46, (note that even in a weighted-average scheme we need a mechanism to update rater credibilities). In the proposed sCARE approach, we start by using the majority function (Equation 3), which yields a majority rating of 0.5. The next step is computing d (Equation 2) for each r_j. The resulting vector is {0, 0.3, 0, 0.3, 0, 0, 0.1, 0.2, 0.2, 0.1}. Assume that x has set a threshold (θ) of 0.1, i.e., if d is greater than 0.1, the credibility of r_j is decreased, and increased otherwise (Equation 7). The resulting credibility vector is then {0.66, 0.49, 0.55, 0.08, 0.33, 0.55, 0.37, 0.44, 0.31, 0.33}. Note that here we assume that the transaction impact factor (τ) is set to 1, i.e., each transaction is equally important. Moreover, Equation 8 is not applied to dilute the rater credibilities. The next step is using the reputation function, i.e., calculate Ex, En, and He. Table I shows the intermediate calculations, and the resulting values are 0.48, 0.13, and 0.06. Since He is less than θ, we choose $Trust(s_i)$ as Ex (from the Reputation Estimation equation above), i.e. 0.48. Note that this value is closer to the majority rating, considering all rater credibilities (updating them as well) in comparison with a simple weighted average-scheme. A detailed experimental study is presented next to show how reputations and credibilities evolve under sCARE.

Table I. Sample Calculations

r_j	d	w^t	w^{t+1}	$w^{t+1}r_j$	$w^{t+1}\lvert r_j - Ex\rvert$	$w^{t+1}(r_j - Ex)^2$
0.5	0	0.6	0.66	0.33	0.02	0.000594
0.2	0.3	0.6	0.49	0.1	0.13	0.035721
0.5	0	0.5	0.55	0.28	0.02	0.000495
0.2	0.3	0.1	0.08	0.02	0.02	0.005832
0.5	0	0.3	0.33	0.17	0.01	0.000297
0.5	0	0.5	0.55	0.28	0.02	0.000495
0.6	0.1	0.4	0.37	0.23	0.05	0.006253
0.7	0.2	0.5	0.44	0.31	0.1	0.023276
0.3	0.2	0.35	0.31	0.1	0.05	0.008959
0.4	0.1	0.35	0.33	0.13	0.02	0.001617
$v = 0.5$		4.2	4.11	1.96	0.44	0.08
	$Ex = 0.48$		$En = 0.13$		$He = 0.06$	

6. EXPERIMENTS

In this section, we provide experiment results to indicate the robustness of the proposed model (sCARE). We show the performance of sCARE in estimating the reputation of different service provider behaviors given variable honest and dishonest ratings. We also compare the overall performance of sCARE, with two similar works. The experiment environment (which mimics real-world service Web interactions) is controlled in the sense that we can monitor the genuine QoS delivered by the service providers, and hence can evaluate their 'actual' trust values. It is assumed that in each time instance, service consumers and providers interact with each other and at the end of the interaction, service consumers rate the service providers. Therefore, the fader value (f_d) is set to 1. For experimental and data gathering purposes, we assume that each service provider offers a single operation that can be invoked by the service consumers. As mentioned previously, the sCARE model (and similar works) estimates the service provider reputations according to the ratings provided by the service raters. The performance of the reputation estimation model is then assessed as the variation between the *actual* and the calculated reputation. Although system performance related to reputation storage and collection overheads is important, it is out of the scope of this work.

Setup: The experiment service Web environment consists of hundred (100) Web services. One round of provider-consumer interactions completes in 1000 time-instances.

For data gathering purposes, we execute fifteen of these rounds, and report the averages in the following. We have used five QoS attributes in the experiments, namely: encryption, availability, invocation fee, response time and authentication (since using more than one criterion provides a better evaluation of the service provider). The values to these attributes are assigned according to a pre-defined rating scheme. We used the the WSDream QoS-Dataset [Zheng and Lyu 2010] to model the different service quality behaviors. This data-set contains around 150 Web services distributed in computer nodes located all over the world (i.e., distributed in 22 different countries), where each Web service is invoked 100 times by a service user. Planet-Lab is employed for monitoring the Web services. The service users observe, collect, and contribute the QoS data of the selected services that is used in modeling rater credibilities, and provider quality patterns.

Distribution of Raters: Analogous to the real-world, a rater is classified as either being honest or dishonest. The classification is performed according to their ratings reporting accuracy. For instance, honest raters tend to provide the true/actual rating for the experienced QoS, while a dishonest rater deviates from the actual rating by at least 0.3. Say the provider's trust value was 0.9, then a dishonest rater would generate a value between [0.1 and 0.59]. The rater class (honest vs. dishonest) is chosen using a capped Gaussian distribution in our experiments. Moreover, there are primarily three ways in which the honest and dishonest raters can co-exist: Honest raters can form the majority, equal proportion of honest and dishonest raters can be present, or dishonest raters can form the majority. We choose a significant majority (80% vs. 20% raters) for case one and three.

Provider Strategies: To simulate real-world behaviors, we divide the service providers into five groups. Each group consists of twenty members (100 total as mentioned above). Service providers in the first group perform with consistently high QoS values, while the second group performs in an opposite manner, i.e., with consistently low QoS values. The third group consists of strategic providers that attempt to take advantage of the consumers by providing high QoS in the beginning, but around the 500th. time instance they start to provide lower QoS values. As mentioned in [Xiong and Liu 2004], this behavior is termed as reputation "milking", where the aim is to build a good reputation in the beginning and then take advantage of the established goodwill. The next group (four) of providers are those that perform in a manner opposite to the third group, i.e. with initially low QoS, after 500 time instances their performance improves. These providers acknowledge their low performance (which may be due to a number of factors in the real-world), and act upon it towards improvement. The fifth provider group performs with random QoS values, i.e., these providers do not follow any given pattern. In one times instance, their QoS may be high, and in the next it may either improve or degrade. The above mentioned groups of providers and combinations thereof, envelop any provider delivered QoS behavior, and are hence representative of the real-world.

sCARE Performance with Variable Rater and Provider Behaviors

Honest Rater Majority: In the first experiment, we set 80% of the raters to behave honestly, i.e., their credibilities are constantly improving, while the rest 20% of the raters should have their credibilities reduced (on account of being dishonest). Figure 1 shows the effect of this inequality in estimating the provider's reputation for the five provider behaviors defined above (labeled A through E). For instance, Figure 1-C shows the case of a provider that performs with high QoS values (denoted original performance in the graphs) for some initial iterations but around the 480-500 iteration mark, the performance degrades. We can see that the estimated reputation values for the five provider behaviors are fairly close to the original performance. This is mainly due to the high number of honest ratings, causing the associated credibilities to weigh higher. The minor disparity between the original and the estimated values can be attributed to the differences in opinions of credible raters, and naturally the 20% offset of dishonest raters (though weighed very lightly).

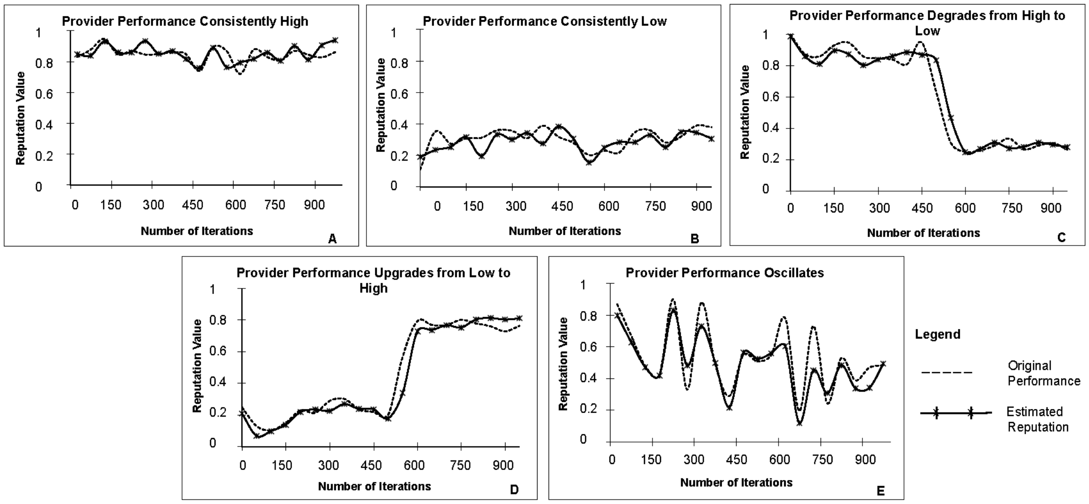

Fig. 1. Majority of Raters have High Credibility

<u>*Honest and Dishonest Rater Equality*</u>: In the second set of experiments, we set half of the raters (i.e. 50%) as honest, and the remaining ones as dishonest. Figure 2 shows that in comparison with the previous case (80-20 imbalance), the estimated reputations exhibit a greater divergence, though still close to the original. Note that for estimation purposes, some variation below a pre-defined threshold is usually acceptable (i.e., still considered accurate). Since majority ratings (v) that are calculated from the reported consumer ratings depend on the generated data clusters, we use a two-point threshold in our experiments, whereby the ratings that differ by two points are allocated to the same cluster. The centroid of a cluster holding maximum elements is the majority rating for a particular time instance. Note that if two or more clusters have equal number of members, then the tie is broken randomly for the choice of the majority rating. Thus, as shown in Figure 2 v (and ultimately the estimated reputation) is sometimes closer to the actual performance, while at others it is not. In the former case, honest raters' cluster centroid is chosen, while in the latter case, dishonest raters' cluster centroid forms the majority rating v.

Dishonest Rater Majority: The third experiment is the opposite of the first, where we set 80% of the raters to behave dishonestly, while the rest 20% of the raters provide honest feedbacks. In Figure 3 we can see that estimated reputations are much closer to the original performance, in comparison with Figure 2 (where a 50-50 balance is maintained between honest and dishonest raters). Note that in the current set of experiments, collusion among raters to elevate or degrade a provider's reputation is not factored, and agreement in ratings is random. Moreover, due to the dishonest rater

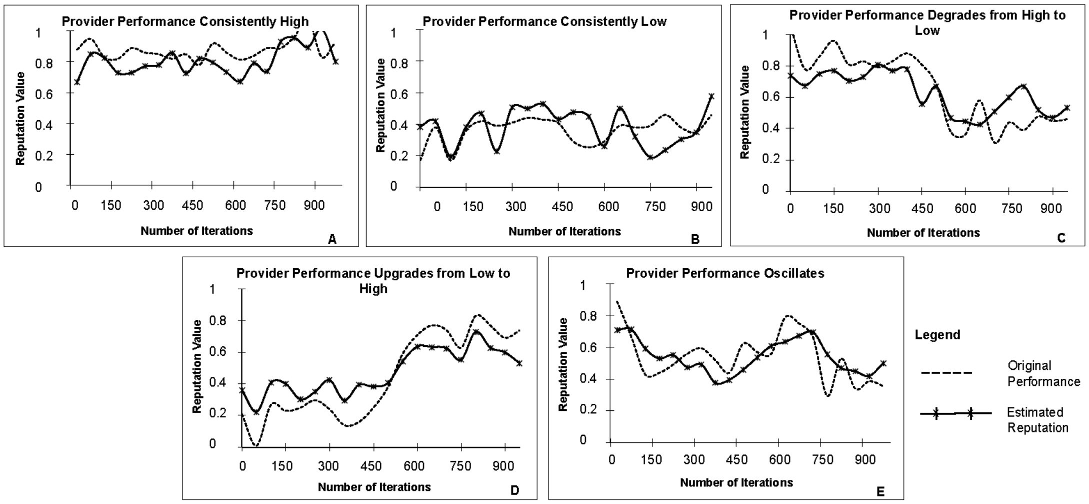

Fig. 2. Equal Number of Raters with High and Low Credibilities

majority, there is some discrepancy between the actual vs. estimated values. However, since the deviation is not large, the estimated values are fairly close. The dishonest majority causes v to 'side' with the dishonest reporting in each instance. Still, as stated in Equation 8, the credibilities are 'kept in check' after each iteration/transaction. Hence, the low disparity among original and estimated values (despite the large number of dishonest ratings). Whitby et al. have reported that having more than 40% of the raters perform in a dishonest manner is unrealistic for real-world applications [Whitby et al. 2005]. Thus, we may safely conclude that sCARE provides a fairly accurate estimation of provider reputations.

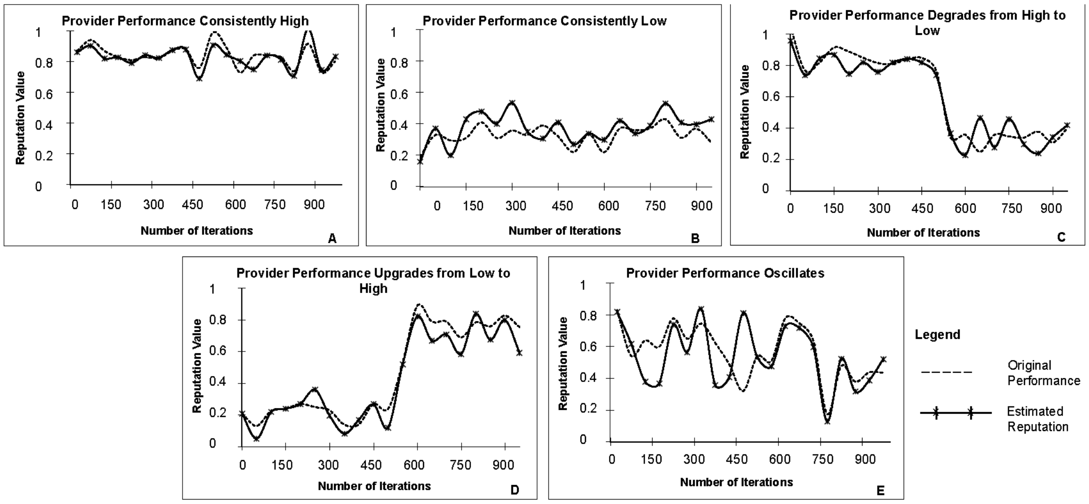

Fig. 3. Majority of Raters have Low Credibility

sCARE Performance in Comparison with Similar Works

In this section, we compare sCARE with a conventional approach where reputations are calculated as simple averages of all the submitted ratings, a variant of a popular heuristics-based approach for P2P systems that also considers rater credibilities (PeerTrust [Xiong and Liu 2004]), and a similar Web services-based reputation system in RATEWeb [Malik and Bouguettaya 2009b]. Like these systems, we distinguish between two types of rater behaviors, without and with collusion. In the former case, service providers perform with low QoS values, and raters provide dishonest ratings, i.e., submit corresponding higher values. In the latter case, service provider behaviors stay the same while service raters collude with other services to increase/decrease some specific provider's reputation. In the following, we refer to dishonest raters as malicious. For experimental purposes, we change the percentage of malicious raters in steps of 10%. At the end of a transaction, if the QoS delivered by the service provider and the estimated one are 'close' (within a pre-determined threshold), then the estimation is considered as successful. Thus, reputation estimation success rate (ESR) is defined as the total number of successful reputation estimations over the total number of reputation estimations in the running environment. From Figure 4 we can see that the ESR drops almost consistently. The primary reason for this output is raters' behaviors, i.e., providing dishonest ratings. As expected, the normal averaging model performs the worst. Of the other three systems, sCARE outperforms both PeerTrust-V and RATEWeb. In both settings, sCARE's ESR drops near that of RATEWeb and PeerTrust-V when 65%-70% of the raters provide dishonest ratings. When majority of the ratings are dishonest, it becomes difficult for the system to assess the"true" reputation. Incorrect (majority) ratings are considered credible in each time instance and ESR drops. PeerTrust-V is best capable to handle collusion in cases of higher percentages of dishonest raters. In the non-collusive setting, the case is almost similar, with the variance in the three models being very minimal. In multiple studies of eBay's reputation system, it has been shown that raters usually provide positive feedbacks for the sellers, and perform honestly [Resnick and Zeckhauser 2002]. This is likely due to the fact that both parties rate each other [Josang et al. 2007]. However, we believe that it provides a rough guideline for the number of credible raters in a ratings-based reputation community. Moreover, it is expected that such high numbers of malicious raters in real world applications are unrealistic and a much lower rate of dishonesty should be expected [Whitby et al. 2005]. Thus, we may conclude that in terms of effectiveness for real world applications, from among the models evaluated, sCARE provides the best level of accuracy.

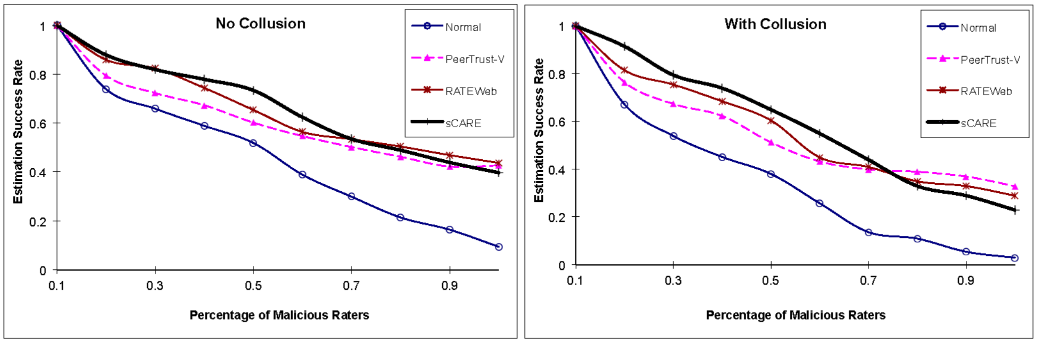

Fig. 4. Measuring Reputation Estimation Success Against Varying Percentage of Dishonest Raters

7. RELATED WORK

Trust assessment involves several components, including modeling, data collection, data storage, communication, assessment, and safeguards. Over the years, varied disciplines including economics, computer science, marketing, politics, sociology, and psychology have studied reputation-based trust in several contexts [Dellarocas 2003]. In the recent past, these research activities have gained momentum. In computer science, reputation has been studied both in theoretical areas and practical applications. Theoretical areas where reputation has been studied include game theory [Ely and Valimaki 2003], Bayesian networks [Wang and Vassileva 2003], overlay networks,[Rocha et al. 2006] and social networks [Buskens 1998] to name a few. Theoretical literature that addressed reputation focused on proving properties of systems based on reputation. For example, results from game theory demonstrate that there are inherent limitations to the effectiveness of reputation systems when participants are allowed to start over with new names [Resnick et al. 2000]. In [Huberman and Wu 2003], the authors study the dynamics of reputation, i.e., growth, decay, oscillation, and equilibria. Practical literature on reputation is mainly concerned with the applications of reputations. Major applications where reputation has been effectively used include e-business, peer-to-peer (P2P) networks, grid computing systems [Azzedin and Maheswaran 2002], multi-agent systems [Sabater and Sierra 2003], Web search engines, and ad-hoc network routing [Boudec 2002]. In the following, we give a brief overview of a few reputation management frameworks for P2P systems and Web services since these are closely related to our research.

BRS [Whitby et al. 2005], BLADE [Regan et al. 2006], and TRAVOS [Teacy et al. 2005] are examples of systems that use Bayesian models to assess the reputations of service providers. The underlying technique is to use the number of satisfactory and unsatisfactory interactions with the service providers and categorize them as ratings. Then, based on statistical analysis (e.g., outlier detection) dishonest ratings are filtered out. A primary drawback of such approaches is that a large amount of relevant information may be categorized as dishonest, and hence discarded. Moreover, these models (e.g., TRAVOS) assume that service providers behave consistently, which might not be the case in real-world scenarios. However, unlike BRS and TRAVOS, BLADE presents a Bayesian model in which dishonest ratings are still considered through a learning approach (similar to ours), in which raters that provide ratings similar to the service consumer's own experience are weighed higher. This implies that BLADE works well in situations where raters are "extreme", i.e. very honest or very dishonest. GMGC is a similar model developed by Chen and Singh [Chen and Singh 2001] which computes rater credibilities. The approach first calculates quality and confidence values of each rating given by the rater to an object. It then computes the quality and confidence values of all ratings for the objects in each category or subcategory. This may prove to be a major limitation in systems with complex object categorization in terms of computation time. The Probabilistic Reputation (PRep) model [Haghpanah and desJardins 2012] is another probabilistic model based on Bayesian learning. In PRep, an agent gathers information about a target agent through both direct interactions, and reviewer ratings. It then learns the reporting agent's behavior by comparing these two sources . The PRep agent can then interpret other ratings for other agents coming from the same rater. A primary limitation of this approach is the requirement of similar rated subjects, which is unlikely in a SOA.

The effects of applying reputation in using game theoretic approaches has also been studied. For example, Harsanyi Agents Pursuing Trust in Integrity and Competence (HAPTIC) [Smith and desJardins 2009], is game theory and probabilistic modeling based framework in which agents learn other agents' behaviors using direct experiences. HAPTIC models trust using two components: competence and integrity. A HAPTIC agent observes the behavior of other agents and estimates their competence and integrity. This information is then used for interacting with other agents. A major limitation of HAPTIC is that it does not support reported experiences. Another typical problem where game theory was instrumental is the *entrance deterrence problem*, also known as Selten's chain store game [Selten 1978]. In this game, a multi-market monopolist faces a finite sequence of potential entrants. Each entrant observes the actions taken in the previous markets and chooses whether to enter the market monopolized by the incumbent firm (based on reputation). Another reputation-related issue studied in game theory is the effect of changing identities on reputation systems. Results from game theory demonstrate that there are inherent limitations to the effectiveness of reputation systems when participants are allowed to start over with new names [Resnick et al. 2000].

PeerTrust [Xiong and Liu 2004] is a P2P reputation management framework used to quantify and compare the trustworthiness of peers. In PeerTrust, the authors have proposed to decouple feedback trust from service trust, which is similar to the approach undertaken in this paper. Similarly, it is argued that peers use a similarity measure to weigh opinions of those peers highly who have provided similar ratings for a common set of past partners. However, this may not be feasible for large P2P systems, where finding a statistically significant set of such past partners is likely to be difficult. In [Kamvar et al. 2003], the *EigenTrust* system is presented, which computes and publishes a global reputation rating for each node in a network using an algorithm similar to Google's *PageRank* [Page et al. 1998]. EigenTrust centers around the notion of transitive trust, where feedback trust and service trust are coupled together. Peers that are deemed honest in resource sharing are also considered credible sources of ratings information. This is in contrast with our approach and we feel this approach may not be accurate. Moreover, the proposed algorithm is complex and assumes existence of pre-trusted peers in the network. PowerTrust [Zhou and Hwang 2007] is a "distributed version" of EigenTrust. It states that the relationship between users and feedbacks on eBay follow a Power-law distribution. It exploits the observation that most feedback comes from few "power" nodes to construct a robust and scalable trust modeling scheme. In PowerTrust, nodes rate each interaction and compute local trust values. These values are then aggregated to evaluate global trust through random walks in the system. Once power nodes are identified, these are used in a subsequent look-ahead random walk that is based on Markov chain to update the global trust values. Power nodes are used to assess the reputation of providers in a "system-wide absolute" manner. This is in contrast with our approach where each consumer maintains control over the aggregation of ratings to define a provider's reputation. Moreover, PowerTrust requires a structured overlay (for DHT), and the algorithms are dependent on this architecture. In contrast, service-oriented environments or the Web in general do not exhibit such structure.

In [Maximillien and Singh 2002], a distributed model for Web service reputation is presented. The model enables a service's clients to use their past interactions with that service to improve future decisions. The authors present an approach that provides a conceptual model for reputation that captures the semantics of attributes. A similar reputation-based model using a node's first hand interaction experience is presented in [Rocha et al. 2006]. The goal of the model is to increase/maintain QoS values in *selfish* overlay networks. The authors show that in presence of a reputation management system, an overlay network discourages *selfish* nodes. This increases the QoS guarantees in the network. The proposed model considers a node's first hand interaction experience and peer testimonials for deriving node reputations. In this regard, the reputation building process in [Rocha et al. 2006] is similar to our approach. However, the proposed reputation model may not be completely robust and may not provide accurate results. First, the individual experience takes time to evolve over repeated interactions. Second, no distinction is made between the node's service credibility in satisfying consumer requests and its rating credibility. It may be the case that a node performs satisfactorily but does not provide authentic testimonials. We provide an extensive mechanism to overcome these and similar inadequacies.

A reputation-based trust mechanism for cloud environments is presented in [Wu 2012]. It uses fuzzy logic to handle uncertain and incomplete trust reports on the cloud. The way this approach deals with uncertainty is limited to using a fuzzy rules table (e.g., "good", "low") to decide about the cloud service quality. [Lu et al. 2014] defines a trust model based on two types of quality of service parameters: objective parameters such as execution time and subjective parameters such as qualitative parameters given by service providers and consumers. Hypothesis testing is used to remove outliers from objective parameters. Subjective parameters are analyzed based on direct and recommended trust. However, [Lu et al. 2014] does not deal with key issues such as the credibility and reputation of providers and consumers when dealing with subjective parameters. [Wang et al. 2011] uses the cloud statistical model to deal with quality of service uncertainty. [Chen et al. 2013] uses collaborative filtering to predict current quality of service values based on the past quality of service evaluations from users. However, both [Wang et al. 2011] and [Chen et al. 2013] use uncertainty for service selection *not* for reputation and trust establishment.

8. CONCLUSION

We have presented a trust estimation model (sCARE) for services-based environments. The proposed model incorporates the uncertainty and fuzziness of trust to provide a more holistic assessment. Our solution is extensible and can be deployed in other contexts and can integrate various functions seamlessly (e.g., the credibility function can be replaced if a better alternative is found). We have focused on a Peer-to-Peer (P2P) environment where Web services can act as both consumers (i.e., requesters) and providers of services without the need of a *trusted third party*. We have also conducted extensive experiments to cover a number of real-world scenarios, and to verify the proposed model. Results from the experiments exhibit strong evidence that our approach estimates provider trust in a fairly accurate manner. A comparative study with similar prior works is also presented to validate sCARE's applicability.

REFERENCES

F. Azzedin and M. Maheswaran. 2002. Evolving and Managing Trust in Grid Cmputing Systems. In *Proc. of the IEEE Canadian Conference on Electrical and Computer Engineering*. 1424–1429.

M. Barhamgi, D. Benslimane, and A. M. Ouksel. 2008. Composing and optimizing data providing web services. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*. 1141–1142.

J. Ben-Naim and H. Prade. 2012. Evaluating trustworthiness from past performances: interval-based approaches. *Ann. Math. Artif. Intell.* 64, 2-3 (2012), 247–268.

K. Benouaret, D. Benslimane, A. Hadjali, M. Barhamgi, Z. Maamar, and Q. Z. Sheng. 2014. Web Service Compositions with Fuzzy Preferences: A Graded Dominance Relationship-Based Approach. *ACM Trans. Internet Techn.* 13, 4 (2014), 12.

E. Bertino, E. Ferrari, and A. C. Squicciarini. 2004. Trust-X: A Peer-to-Peer Framework for Trust Establishment. *IEEE TKDE* 16, 7 (2004), 827–842.

K. Bharadwaj and M. Al-Shamri. 2009. Fuzzy computational models for trust and reputation systems. *Electron. Commer. Rec. Appl.* 8, 1 (2009), 37–47. DOI:http://dx.doi.org/10.1016/j.elerap.2008.08.001

K. Birman. 2006. The Untrustworthy Web Services Revolution. *IEEE Computer* 39, 2 (2006), 113–115.

S. Buchegger J.-Y. Le Boudec. 2002. Performance Analysis of the CONFIDANT Protocol. In *Proc. of the 3rd ACM Intl. Symposium on Mobile Ad Hoc Networking and Computing*. 226–236.

S. Buchegger and J.-Y. Le Boudec. 2004. A Robust Reputation System for P2P and Mobile Ad-hoc Networks. In *Proceedings of the Second Workshop on the Economics of Peer-to-Peer Systems*.

V. Buskens. 1998. Social Networks and the Effect of Reputation on Cooperation. In *Proc. of the 6th Intl. Conf. on Social Dilemmas*.

M. Chen and J. P. Singh. 2001. Computing and Using Reputations for Internet Ratings. In *Proceedings of the 3rd ACM Conference on Electronic Commerce (EC '01)*. 154–162.

X. Chen, Z. Zheng, X. Liu, Z. Huang, and H. Sun. 2013. Personalized QoS-Aware Web Service Recommendation and Visualization. *IEEE T. Services Computing* 6, 1 (2013), 35–47.

J. Delgado and N. Ishii. 1999. Memory-Based Weighted-Majority Prediction for Recommender Systems. In *ACM Workshop on Recommender Systems*.

C. Dellarocas. 2003. The Digitalization of Word-of-Mouth: Promise and Challeges of Online Feedback Mechanisms. *Management Science* (October 2003).

J. C. Ely and J. Valimaki. 2003. Bad Reputation. *Quarterly Journal of Economics* 118 (2003), 785–814.

Y. Haghpanah and M. desJardins. 2012. PRep: A Probabilistic Reputation Model for Biased Societies. In *11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1 (AAMAS '12)*. 315–322.

R. He, J. Niu, M. Yuan, and J. Hu. 2004. A Novel Cloud-Based Trust Model for Pervasive Computing. *Computer and Information Technology, International Conference on* 0 (2004), 693–700.

D. Houser and J. Wooders. 2005. Reputation in Auctions: Theory, and Evidence from eBay. *Journal of Economics and Management Strategy* (2005).

B. A. Huberman and F. Wu. 2003. The Dynamics of Reputations. *TR, Hewlett-Packard Laboratories and Stanford University* (January 2003).

A. Jøsang. 2001. A logic for uncertain probabilities. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 9, 3 (2001), 279–311.

A. Josang, R. Ismail, and C. Boyd. 2007. A survey of trust and reputation systems for online service provision. *Decis. Support Syst.* 43, 2 (2007), 618–644. DOI:http://dx.doi.org/10.1016/j.dss.2005.05.019

S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. 2003. The EigenTrust Algorithm for Reputation Management in P2P Networks. In *Proceedings of the Twelfth International World Wide Web Conference (WWW)*.

S. Lam and J. Riedl. 2004. Shilling Recommender Systems for Fun and Profit. In *Proc. of the 13th International World Wide Web Conference (WWW)*. New York, NY, USA, 393–402.

P. Laureti, L. Moret, Yi-Cheng Zhang, and Y.-K. Yu. 2006. Information filtering via Iterative Refinement. *CoRR* abs/physics/0608166 (2006). http://arxiv.org/abs/physics/0608166

D. Li, J. Han, X. Shi, and M.-C. Chan. 1998. Knowledge representation and discovery based on linguistic atoms. *Knowledge-Based Systems* 10, 7 (1998), 431 – 440.

D. Li, C. Liu, and W. Gan. 2009. A new cognitive model: Cloud model. *Int. J. Intell. Syst.* 24, 3 (2009), 357–375.

W. Lu, X. Hu, S. Wang, and Xiaotao Li1. 2014. A Multi-Criteria QoS-aware Trust Service Composition Algorithm in Cloud Computing Environments. *International Journal of Grid and Distributed Computing* 7, 1 (2014), 77–88.

R. Malaga. 2001. Web-Based Reputation Management Systems: Problems and Suggested Solutions. *Electronic Commerce Research* 1, 1 (2001), 403–417.

Z. Malik and A. Bouguettaya. 2009a. Rater Credibility Assessment in Web Services Interactions. *World Wide Web Journal* 12, 1 (March 2009), 3–25.

Z. Malik and A. Bouguettaya. 2009b. RATEWeb: Reputation Assessment for Trust Establishment Among Web Services. *The VLDB Journal* 18, 4 (Aug. 2009), 885–911.

Z. Malik and A. Bouguettaya. 2009c. Reputation Bootstrapping for Trust Establishment among Web Services. *IEEE Internet Computing* 13, 1 (January-February 2009).

Z. Malik and A. Bouguettaya. 2009d. *Trust Management for Service-Oriented Environments* (1 ed.). Springer. ISBN:978-1-4419-0309-9.

S. Marti and H. Garcia-Molina. 2004. Limited Reputation Sharing in P2P Systems. In *Proc. of the 5th ACM Conference on Electronic Commerce*. New York, NY, USA, 91–101.

E. M. Maximillien and M.P. Singh. 2002. Conceptual Model of Web Service Reputation. *SIGMOD Record* 31, 4 (December 2002), 36–41.

B. Medjahed and A. Bouguettaya. 2005. Customized Delivery of E-Government Web Services. *IEEE Intelligent Systems* 20, 6 (November/December 2005).

B. Medjahed, A. Bouguettaya, and A. Elmagarmid. 2003. Composing Web Services on the Semantic Web. *The VLDB Journal* 12, 4 (November 2003).

J. Niu, Z. Chen, and G. Zhang. 2006. Towards A Subjective Trust Model with Uncertainty for Open Network. *Grid and Cooperative Computing Workshops, International Conference on* 0 (2006), 102–019.

T. H. Noor, Q. Z. Sheng, and Athman Bouguettaya. 2014. *Trust Management in Cloud Services*. Springer.

T. H. Noor, Q. Z. Sheng, S. Zeadally, and J. Yu. 2013. Trust management of services in cloud environments: Obstacles and solutions. *ACM Comput. Surv.* 46, 1 (2013), 12.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford Digital Library Technologies Project.

M.P. Papazoglou and D. Georgakopoulos. 2003. Serive-Oriented Computing. *Communcications of the ACM* 46, 10 (2003), 25–65.

K. Regan, P. Poupart, and R. Cohen. 2006. Bayesian Reputation Modeling in E-marketplaces Sensitive to Subjectivity, Deception and Change. In *21st National Conference on Artificial Intelligence (AAAI'06)*. 1206–1212.

P. Resnick and R. Zeckhauser. 2002. *Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System*. Advances in Applied Microeconomics, Vol. 11. Elsevier Science, Amsterdam.

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

B. G. Rocha, V. Almeida, and D. Guedes. 2006. Increasing QoS in Selfish Overlay Networks. *IEEE Internet Computing* 10, 3 (May-June 2006), 24–31.

J. Sabater and C. Sierra. 2003. Bayesian Network-Based Trust Model. In *Proc. of the first Intl. Joint Conf. on Autonomous Agents and Multiagent Systems*. Bologna, Italy, 475 – 482.

R. Selten. 1978. The Chain Store Paradox. *Theory and Decision* 9 (1978), 127–159.

Z. Shibin, S. Xiang, and Q. Zhi. 2009. Subjective Trust Evaluation Model Based on Fuzzy Reasoning. *Electronic Commerce and Security, International Symposium* 1 (2009), 328–332.

M. Smith and M. desJardins. 2009. Learning to trust in the competence and commitment of agents. *Autonomous Agents and Multi-Agent Systems* 18, 1 (2009), 36–82.

W. T. L. Teacy, J. Patel, N. Jennings, and M. Luck. 2005. Coping with Inaccurate Reputation Sources: Experimental Analysis of a Probabilistic Trust Model. In *Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '05)*. 997–1004.

M. Tennenholtz. 2004. Reputation Systems: An Axiomatic Approach. In *AUAI '04: 20th conference on Uncertainty in artificial intelligence*. 544–551.

L.-H. Vu, M. Hauswirth, and K. Aberer. 2005. QoS-based Service Selection and Ranking with Trust and Reputation Management. In *13th International Conference on Cooperative Information Systems (CoopIS 2005)*.

K. Walsh and E. G. Sirer. 2005. Fighting peer-to-peer SPAM and decoys with object reputation. In *P2PECON '05: ACM SIGCOMM workshop on Economics of peer-to-peer systems*. 138–143.

S. Wang, Z. Zheng, Q. Sun, H. Zou, and F. Yang. 2011. Reliable web service selection via QoS uncertainty computing. *IJWGS* 7, 4 (2011), 410–426.

Y. Wang and J. Vassileva. 2003. Trust and reputation model in peer-to-peer networks. In *Proc. of the Third International Conference on Peer-to-Peer Computing*. 150–158.

J. Weng, C. Miao, and A. Goh. 2005. Protecting Online Rating Systems from Unfair Ratings. *Trust, Privacy and Security in Digital Business* 3592 (August 2005), 50–59.

A. Whitby, A. Josang, and J. Indulska. 2005. Filtering Out Unfair Ratings in Bayesian Reputation Systems. *The Icfain Journal of Management Research* 4, 2 (February 2005), 48–64.

X. Wu. 2012. A Fuzzy Reputation-based Trust Management Scheme for Cloud Computing. *International Journal of Digital Content Technology and its Applications* 6, 17 (2012), 437–445.

L. Xiong and L. Liu. 2004. PeerTrust: Supporting Reputation-based Trust for Peer-to-Peer Electronic Communities. *IEEE Trans. on Knowledge and Data Engineering (TKDE)* 16, 7 (July 2004), 843–857.

S. Zhang, Y. Ouyang, J. Ford, and F. Makedon. 2006. Analysis of a Low-dimensional Linear Model Under Recommendation Attacks. In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. 517–524.

Z. Zheng and M. R. Lyu. 2010. Collaborative Reliability Prediction for Service-Oriented Systems. In *Proc. IEEE/ACM 32nd Int'l Conf. Software Engineering (ICSE'10)*. 35–44.

R. Zhou and K. Hwang. 2007. PowerTrust: A Robust and Scalable Reputation System for Trusted Peer-to-Peer Computing. *IEEE Transactions on Parallel and Distributed Systems* 18, 4 (2007), 460–473.