A Survey of Multi-agent Trust Management Systems

Han Yu[1], Zhiqi Shen[1], Cyril Leung[2], Chunyan Miao[1], and Victor R. Lesser[3]

[1]School of Computer Engineering, Nanyang Technological University, Singapore 639798
[2]Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, V6T1Z4, Canada
[3]School of Computer Science, University of Massachusetts Amherst, Amherst, MA 01003, U.S.

In open and dynamic multi-agent systems (MASs), agents often need to rely on resources or services provided by other agents in order to accomplish their goals. During this process, agents are exposed to the risk of being exploited by others. These risks, if not mitigated, can cause serious breakdowns in the operation of MASs and threaten their long term wellbeing. To protect agents from the uncertainty in the behavior of their interaction partners, the age-old mechanism of trust between human beings has been re-contexted into MASs. The basic idea is to let agents self-police the MAS by rating each other based on their observed behavior and basing future interaction decisions on such information. Over the past decade, a large number of trust management models have been proposed. However, there is a lack of research effort in several key areas which are critical to the success of trust management in MASs where human beings and agents co-exist. The purpose of this article is to give an overview of existing research in trust management in MASs. We analyze existing trust models from a game theoretic perspective to highlight the special implications of including human beings in an MAS, and propose a possible research agenda to advance the state of the art in this field.

***Index Terms*—Trust, reputation, multi-agent systems, game theory**

I. Introduction

MANY systems that are commonplace in our lives, such as e-commerce platforms, crowdsourcing systems, online virtual worlds, and P2P file sharing systems, can be modeled as open dynamic multi-agent systems (MASs). The agents in these systems can represent software entities or human beings. They are considered *open* because agents can come from any background with heterogeneous abilities, organizational affiliations, credentials, etc. They are considered *dynamic* because the decision-making processes of the agents are independent from each other and agents can join or leave the system at will. As each agent has only limited capabilities, it may need to rely on the services or resources from other agents in order to accomplish its goals. In these situations, agents often cannot preview quality of the services or resources. Their decisions to trust on another agent involve a certain level of risk. In this relationship, the agent that needs to rely on another agent is referred to as the *truster agent*; the agent that provides resources or services to the truster agent is referred to as the *trustee agent* [1]. These roles tend to be situational rather than fixed since an agent may act as either a truster or a trustee under different circumstances.

Trust was first introduced as a measurable property of an entity in computer science in [65]. Following this work, a number of computational models focusing on various facets of trust management have been proposed in the MAS literature. In general, the act of trusting by an agent can be conceptualized as consisting of two main steps: 1) *trust evaluation*, in which the agent assesses the trustworthiness of potential interaction partners; and 2) *trust-aware decision-making*, in which the agent selects interaction partners based on their trust values. Table I lists research papers on trust management published during the period 2002 to 2012. As shown in Table I, the majority of the research effort is currently focused on the trust evaluation sub-field. This sub-field deals with the problem of accurately evaluating the trustworthiness of potential interaction partners. The proposed methods can be divided into four main categories, 1) *direct trust evaluation models*, which rely on past observed behaviors; 2) *indirect/reputation-based trust evaluation models*, which rely on third-party testimonies from other agents in the same environment; 3) *socio-cognitive trust evaluation models*, which rely on analyzing the social relationships among agents to estimate their trustworthiness; and 4) *organizational trust evaluation models*, which rely on the organizational affiliation or certificates issued by some trusted organizations to estimate the trustworthiness of agents. Compared with the trust evaluation sub-field, very limited research has been done in the trust-aware interaction decision-making sub-field.

Apart from these two major research sub-fields, assessing the performance of proposed agent trust models is also an important sub-field in agent trust research. Although datasets concerning certain aspects of the trust evaluation problem are available (e.g., the *Epinions* and Extended *Epinions* datasets [27]), it is often difficult to find suitable real world data since the effectiveness of various trust models need to be assessed under different environmental conditions and misbehaviors. Therefore, in the current agent trust research field, most of the existing trust models are assessed using simulation or synthetic data. One of the most popular simulation test-beds for trust models is the agent reputation and trust (ART) test-bed proposed in [6]. However, even this test-bed does not claim to be able to simulate all experimental conditions of interest. For this reason, many researchers design their own simulation environments when assessing the performance of their proposed trust models.

In the current multi-agent trust research landscape, agents are normally considered to be *selfish* - meaning that an agent will take whatever action that is expected to maximize its own

TABLE I
SUMMARY OF MULTI-AGENT TRUST RESEARCH PAPERS PUBLISHED IN AAAI, AAMAS AND IJCAI (2002-2012)

Direct Trust	Indirect Trust	Socio-cognitive Trust	Organizational Trust	Trust-aware Interaction Decision-making	Performance Assessment
Tran and Cohen [2]	Tran [3]	Castelfranchi *et al.* [4]	Kollingbaum and Norman [7]	Fullam and Barber [5]	Fullam *et al.* [6]
Griffiths [8]	Yu and Singh [9]	Falcone and Castelfranchi [13]	Huynh *et al.* [10]	Teacy *et al.* [11]	Kerr and Cohen [12]
Zheng *et al.* [38]	Teacy *et al.* [14]	Falcone *et al.* [15]	Hermoso *et al.* [16]	Burnett *et al.* [17]	
Bedi *et al.* [18]	Regan *et al.* [19]	Ashri *et al.* [20]		Pasterneck and Roth [21]	
Dondio and Barrett [23]	Wang and Singh [22]	Casare and Sichman [24]			
Osman and Robertson [25]	Fullam and Barber [26]	Messa and Avesani [27]			
Reece *et al.* [28]	Hendrix and Grosz [29]	Katz and Golbeck [30]			
Wang and Singh [32]	Kawamura *et al.* [31]	Kuter and Golbeck [33]			
Reches *et al.* [34]	Procaccia *et al.* [35]	O'Donovan *et al.* [36]			
Teacy *et al.* [37]	Wang and Singh [32]	Rettinger *et al.* [38]			
Khosravifar *et al.* [41]	Hang *et al.* [39]	Burnett *et al.* [40]			
Matt *et al.* [42]	Tang *et al.* [43]	Koster *et al.* [44]			
Salihi-Abari and White [47]	Liu *et al.* [45]	Li and Wang [46]			
Vogiatzis *et al.* [48]	Zhang *et al.* [49]	Liu *et al.* [50]			
Koster *et al.* [44]	Fang *et al.* [51]	Liu and Datta [52]			
Pasternack and Roth [21]	Haghpanah and desJar-dins [54]	Noorian *et al.* [53]			
Witkowski [55]	Koster *et al.* [56]	Singh [57]			
Burnett and Oren [60]	Liu *et al.* [58]	Venanzi *et al.* [59]			
Jiang *et al.* [61]	Piunti *et al.* [62]	Liu and Datta [63]			
	Serrano *et al.* [64]				

utility. Although there has been some preliminary attempts at studying the influence of irrational behaviors (e.g., emotion) on trust among agents [66], irrational agents are usually not among the primary focuses of research in MAS.

The goal of trust-aware interaction decision-making is to help a truster agent decide which candidate trustee agent is to be selected to perform a given task at the cur-rent time. There is a consensus within the current multi-agent trust community that, in order to minimize a truster agent's risk exposure, it should always interact with the trustee agent with the highest reputation that it can find for the given type of task. This approach is a rational choice from the perspective of an individual truster agent and it is adopted by the majority of existing trust models.

However, in a system involving trust-aware interaction decision-making approaches, truster agents are not the only stakeholders. The trustee agent and the collective utility derived through the interactions of all agents in the MAS can also be affected by the interaction decisions made by the truster agents. In principle, trust-aware interaction decision-making approaches should reward trustee agents with high reputations with more tasks so that they can derive more utility through completing them. Over time, it should help the MAS exclude untrustworthy trustee agents and sustain repeated interactions among agents over the long term. However, a closer look at the assumptions used in existing trust models reveals that there are limitations to the applicability of this conclusion.

In this paper, we offer a review of the existing literature in trust management for MASs based on the trust evaluation, interaction decision-making, and performance assessment approaches used in each different trust model. We adopt a game theoretic perspective when viewing the problems that existing trust models are attempting to address and analyze their advantages and potential areas for improvement. The paper is organized as follows. In Section II, the trust games used by many existing trust models to formulate their problems are presented. This is followed by an analysis of the common assumptions made by existing trust models when proposing solutions to these games. Section III reviews the current landscape of multi-agent trust management from the angles of their trust evaluation, interaction decision-making, and performance assessment approaches. In Section IV, we propose an extension to the existing trust game formulations based on our observations. The effectiveness of directly applying existing trust-aware decision-making approaches under the extended trust game is studied in Section V. Section VI presents the potential implications for future trust model designs based on the extended trust game and highlights some open research issues.

II. COMMON ASSUMPTIONS

Trust and reputation modeling in the context of MASs serves the main purpose of forming coalitions for long term interactions among agents who may not know each other at the beginning. During this process, trust among agents acts as a social capital that can affect their future payoffs. Since the act of trusting an agent involves both potential gain and cost for the truster agent, the payoff from trusting is intuitively defined as the difference between these two factors. Under different system models, the derivation of the payoff for the truster agent may be different. This can result in trust building being viewed as a game. In [67], the authors formalized the mechanism of trust in MASs into several types of games:

1) *Additive Trust Game*: this is the base-case game where there is no marginal benefit for the agents to form coalitions through trusting others.
2) *Constant-sum Trust Game*: this is a special case of the additive trust game where the contribution of an agent in a trust relationship is exactly the same as when it is acting alone. Under these two types of system conditions, agents are not motivated to cooperate with each other.
3) *Superadditive Trust Game*: in this case, agents cooperating with others through trust mechanisms derive payoffs which are never less than the sum of payoffs gained through the agents acting alone. The benefit of trusting may theoretically snowball in an MAS, eventually causing all agents to form a grand coalition.
4) *Convex Trust Game*: under this type of game, the benefit for agents to join a coalition increases as the size of the coalition increases. The marginal contribution of agents coming into the coalition is non-decreasing.

The principles established in [67] have been implicitly applied in most of the existing multi-agent trust models. In order to form cooperative relationships based on trust, many assumptions have been made. These assumptions can be classified into two categories:

1) *Fundamental assumptions*: the ones that are essential for multi-agent trust research to be carried out and are commonly accepted by researchers in this field. They include:
   a) Trustee and truster agents are self-interested;
   b) At least one identity can be used to identify a trustee agent;
   c) Every truster agent always prefers to interact with the most trustworthy trustee agent.
2) *Simplifying assumptions*: the ones that are made to enable certain trust models to operate and are not necessarily adopted by many researchers in this field. They include:
   a) The outcome of an interaction between a truster agent and a trustee agent is binary (success or failure);
   b) The effect of an interaction between a truster agent and a trustee agent on the wellbeing of the truster agent can be known immediately after the interaction is completed;
   c) Interactions between a truster agent and a trustee agent occur in discrete time steps;
   d) The majority of third-parties testimonies are reliable;
   e) A truster agent's own direct interaction experience with a trustee agent is the most relevant to itself;
   f) The properties of a trustee agent are useful for predicting its future behavior;
   g) A truster agent needs to select only one trustee agent for each interaction;
   h) A trustee agent can service an unlimited number of requests from truster agents during a time step without affecting its quality of service.

While the fundamental assumptions have stayed the same over the past decade, some of the simplifying assumptions have been relaxed as the characteristics of the application domains evolve. For example, Assumption 2.a was relaxed in [68]; Assumption 2.c was relaxed in [58]; and Assumption 2.d was relaxed in [14]. Based on these assumptions, many multi-agent trust models have been proposed in order to solve the trust games.

III. Trust Models in Multi-agent Systems

Trustworthiness evaluation models employ probabilistic, socio-cognitive, and organizational techniques to enable truster agents to estimate the potential risk of interacting with a given trustee agent. Once the trustworthiness evaluations for a set of candidate trustee agents have been completed, trust-aware interaction decision-making approaches help the truster agent to select a trustee agent for interaction at a particular point in time. By reviewing the key advancements in the multi-agent trust research literature, it can be seen that most existing research is concentrated on improving the accuracy of trust evaluation models. These models can be classified according to their approaches, namely:

1) Direct trust evaluation models,
2) Indirect/Reputation-based trust evaluation models,
3) Socio-cognitive trust evaluation models, and
4) Organizational trust evaluation models.

A. Direct Trust Evaluation Models

One mechanism used by human beings to establish trust between each other is through observing the out-comes of past interactions between them. This evidence based approach of evaluating the trustworthiness of a potential interaction partner has been widely adopted by the multi-agent trust research community. An intuitive way of modeling trust between agents is to view interaction risk as the probability of being cheated by the interaction partner. This probability can be estimated by a truster agent from the outcomes of past interactions with a trustee agent. The historical interaction outcomes serve as the direct evidence available for the truster agent to evaluate a trustee agent's trustworthiness.

One of the earliest models that attempt to derive a trustworthiness value based on direct evidence is the Beta Reputation System (BRS) proposed in [69]. The model, inspired by the Beta Distribution, projects past interaction experience with a trustee agent into the future to give a measure of its trustworthiness. BRS estimates the trustworthiness of a trustee agent by calculating its reputation, which is defined as the probability expectation value of a distribution consists of the positive and negative feedbacks about the trustee agent. This expectation value is then discounted by the belief, disbelief and uncertainty with respect to the truthfulness of the feedbacks (in the case of direct evidence, the truster agent can be certain about the truthfulness since the feedbacks were produced by itself) and then discounted by a forgetting factor to allow past evidence to be gradually discarded. The resulting value is the reputation of the trustee agent.

In BRS, the outcome of an interaction is represented by a binary value (i.e., the interaction is regarded as either a complete success or a complete failure). In order to handle cases where the interaction outcomes are rated on a multinomial scale, Jøsang and Haller introduced the Dirichlet Reputation System (DRS) in [68]. The basic idea behind this model is similar to that in BRS except when modeling the outcomes of historical interactions. However, instead of rating an interaction outcome as a binary value, the outcome of an interaction can take on a value of i where $i = \{1, ..., k\}$ (e.g., a rating of 1 to 5 where 1 represents most unsatisfactory and 5 represents the most satisfactory outcome). With more finely grained ratings, multiple ways of deriving the reputation of a trustee agent are available in DRS. It can be represented as 1) an evidence representation, 2) a density representation, 3) a multinomial probability representation, or 4) a point estimate representation. The first two representations are more difficult for human interpretation than the last two types. In practice, BRS is more widely adopted than DRS.

To gauge the performance of a trustee agent, various aspects of the quality of service provided by it should be analyzed. In [8], a multi-dimensional trust model is proposed that assesses the trustworthiness of a trustee agent along four dimensions: 1) the likelihood that it can successfully produce an interaction result, 2) the likelihood of producing an interaction result within the expected budget, 3) the likelihood of completing the task within the deadline specified, and 4) the likelihood that the quality of the result meets expectation. A weighted average approach is used to compute the trustworthiness of an agent based on these dimensions where the weights are specified by individual truster agents according to their personal preferences.

The work by Wang and Singh in [32] focused on another important aspect of evidence-based trust models - quantifying the uncertainty present in the trust evidence. Consider a scenario where one truster agent A has only interacted with a trustee agent C twice, and in both instances, the outcomes are successful; whereas truster agent B has interacted with C for 100 times and only 50 interactions are successful. Which set of evidence contains more uncertainty for evaluating C's trustworthiness? In [32], this problem was addressed by proposing a method to calculate the uncertainty in a set of trust evidence based on the distribution of positive and negative feedbacks. Based on statistical inference, the method produces a certainty value in the range of [0, 1] where 0 represents the least certainty and 1 represents the most certain. The method satisfies the intuition that 1) certainty is high if the amount of trust evidence is large; and 2) certainty is high if the conflicts among the feedbacks are low.

In practice, the trustworthiness of a trustee agent is often defined within a certain context. This allows individual truster agents to simplify complex decision-making scenarios and focus on the evidence which is most relevant to the interaction decision that has to be made at the moment. Existing evidence-based trust models often handle context by storing past evidence according to the context they belong to. This makes the evaluated trustworthiness valid only within the stipulated context (e.g., a trustee agent's trustworthiness in repairing computers may say little about its trustworthiness in selling T-shirts).

B. Reputation-based Trust Evaluation Models

While direct evidence is one of the most relevant sources of information for a truster agent to evaluate a trustee agent, such information may not always be available. This is especially the case when a large number of agents exist in an MAS and interactions among them are rare. Therefore, indirect evidence (third-party testimonies which are derived from direct interaction experience between a trustee agent and other "witness" agents) may be needed to complement direct evidence for estimating a trustee agent's trustworthiness. However, doing so exposes the truster agents to a new risk - the possibility of receiving biased testimonies which can negatively affect the trust-aware interaction decisions.

The importance of incorporating mechanisms to mitigate the adverse effects of biased testimonies is widely recognized within the research community. In this section, we discuss some recent research work on aggregating trust evidence from different sources and filtering out biased testimonies.

1) Trust Evidence Aggregation Approaches

Evidence-based trust models often make use of two distinct sources of information to evaluate the trustworthiness of a trustee agent: 1) direct trust evidence: a truster agent's personal interaction experience with a trustee agent, and 2) indirect trust evidence: third-party testimonies about the trustee agent. The majority of existing trust models adopt a weighted average approach when aggregating these two sources of trust evidence. Direct trust evidence is often assigned a weight of γ, $(0 \leq \gamma \leq 1)$, and indirect evidence is assigned a corresponding weight of $(1 - \gamma)$. Existing approaches for aggregating direct and indirect trust evidence can be divided into two broad categories: 1) static approaches, where the value of γ is pre-defined; and 2) dynamic approaches, in which the value of γ is continually adjusted by the truster agent.

Static γ values are used in many papers. The majority of them take a balanced approach by assigning a value of 0.5 to γ [70], [71], [45], [72], [73]. In some studies, the authors assign values of 0 [74], [75] or 1 [76] to γ to exclusively use only one source of trust information. Barber and Kim [77] have empirically shown, without considering the presence of biased testimonies, that direct trust evidence is the most useful to a truster agent over the long term while indirect trust evidence gives an accurate picture more quickly. Thus, approaches that discard one source of evidence or the other, forfeit some of the advantages provided by evidence based trust models. Using a static value for γ is generally not a good strategy.

Some researchers have explored methods for adjusting the value of γ dynamically. In [78], the value of γ is varied according to the number of direct observations on the behavior of a trustee agent available to a truster agent. It is assumed that every truster agent starts with no prior interaction experience with a trustee agent and gradually accumulates direct trust evidence over time. Initially, the truster agent relies completely on indirect trust evidence (i.e., $\gamma = 0$) to select trustee agents for interaction. As the number of its interactions with a trustee agent C increases, the value of γ also increases according to the formula

$$\gamma = \begin{cases} \frac{N_C^B}{N_{min}}, \text{ if } N_C^B < N_{min} \\ 1, \text{ otherwise} \end{cases} \quad (1)$$

where N_C^B is the total number of direct observations of a C's behavior by a truster agent B, and N_{min} is the minimum number of direct observations required to achieve a pre-determined acceptable error rate ε and confidence level ϑ. N_{min} is calculated from the *Chernoff Bound Theorem* as:

$$N_{min} = -\frac{1}{2\varepsilon^2} ln \frac{1-\vartheta}{2}. \quad (2)$$

This approach is not concerned with filtering potentially biased third-party testimonies. Rather, its aim is to accumulate enough direct trust evidence so that a truster agent can make a statistically accurate estimate of the trustworthiness of a trustee agent without relying on indirect trust evidence. In order to achieve a high level of confidence and a low error rate, N_{min} may be very high. In practice, this may mean a significant risk to the truster agent. Moreover, since the value of γ increases to 1, this approach implicitly assumes that agent behaviors do not change with time. This may not always be true and limits the applicability of the approach under more dynamic scenarios.

In [26], an approach based on the Q-learning technique [79] is proposed to select an appropriate γ value from a predetermined static set, Γ, of values. In order to select appropriate values for Γ, expert opinions about the underlying system characteristics are assumed to be available. Based on the reward accumulated by a truster agent under different γ values, Q-learning selects the γ value associated with the highest accumulated reward at each time step. This work provided the first step towards using interaction outcomes to enable the truster agent to weight the two sources of trust evidence. However, as this method uses a predetermined set of γ values, its performance is affected by the quality of the expert opinions used to form the set of permissible γ values.

2) Testimony Filtering Approaches

Over the years, many models for filtering potentially biased third-party testimonies have been proposed. However, these models usually assume the presence of some infrastructure support or special characteristics in the environment. In this section, some representative models in this sub-field are discussed.

The ReGreT model [80] makes use of the social relationships among the members of a community to deter-mine the credibility of witnesses. Pre-determined fuzzy rules are used to estimate the credibility of each witness which is then used as the weight of its testimony for a trustee agent when aggregating all the testimonies. This model relies on the availability of social network information among the agents which may not be available in many systems.

In [81], unfair testimonies are assumed to exhibit certain characteristics. The proposed approach is closely coupled with the Beta Reputation System [69] which records testimonies in the form of counts of successful and unsuccessful interactions with a trustee agent. The received testimonies are aggregated with equal weights to form a majority opinion and then, each testimony is tested to see if it is outside the q quartile and $(1-q)$ quartile of the majority opinion. If so, the testimony is discarded and the majority opinion updated. This model assumes that the majority opinion is always correct. Thus, it is not effective in highly hostile environments where the majority of witnesses are malicious.

In [70], it is assumed that the direct experience of the truster agent is the most reliable source of belief about the trustworthiness of a particular trustee agent, and it is used as the basis for filtering testimonies before aggregating them to form a reputation evaluation. An entropy-based approach is proposed to measure how much a testimony deviates from the current belief of the truster agent before deciding whether to incorporate it into the current belief. However, by depending on having sufficient direct interaction experience with a trustee agent, this assumption conflicts with the purpose for relying on third-party testimonies, which is to help truster agents make better interaction decisions when they lack direct trust evidence.

The temporal aspect of the behavior data of witnesses is studied in [82] and a model for filtering potentially unfair testimonies is proposed. The authors designed an online competition platform to let test users deliberately attack it by giving out biased ratings for virtual products. The proposed model - TAUCA - combines a variant of the cumulative sum (CUSUM) approach [83] that identifies the point in time when possible changes in witness behavior patterns occur with correlation analysis to filter out suspicious testimonies. The model has been shown to be robust against Sybil attacks where an attacker controls multiple identities and uses them to give out unfair ratings.

The model in [45] supports interaction outcomes recorded in multi-dimensional forms. It applies two rounds of clustering of the received testimonies to identify testimonies which are extremely positive or extremely negative about a trustee. If neither the extremely positive opinion cluster nor the extremely negative opinion cluster forms a clear majority, they are both discarded as unfair testimonies and the remaining testimonies are used to estimate the reputation of a trustee agent. Otherwise, the majority cluster is considered as the reliable testimonies. Due to its iterative nature, the computational complexity of this method is high, with a time complexity of $O(mn^2)$ where m is the number of candidate trustee agents whose reputations need to be evaluated and n is the number of testimonies received for each candidate trustee agent. The method is also not robust in hostile environments where the majority of the witnesses are malicious.

The effectiveness of many existing reputation-based trust evaluation models depends on witness agents sharing their prior experience interacting with a trustee agent all at once.

However, in practice, such information is often obtained piecemeal, and thus, requires to be maintained over time. The approach of discounting past trust evidence through a temporal discount factor is widely used [69], [84]. In [85], a mechanism enabling a truster agent to update its trust on a trustee agent on an ongoing basis is proposed. It considers trust and certainty together and allows both measures to vary incrementally when new evidence is made available. The mechanism also provides a way for the truster agent using it to avoid the need to require human intervention for parameter tuning. In this way, there is less uncertainty in the performance of the proposed mechanism.

3) Socio-cognitive Trust Evaluation Models

Another school of thought in multi-agent trust re-search emphasizes the analysis of the intrinsic properties of the trustee agents and the external factors affecting the agents to infer their likely behavior in future interactions. This category of trust models are mainly designed to complement evidence-based trust models in situations where there is not enough evidence to draw upon when making trusting decisions.

In [4], a trust decision model based on the concept of fuzzy cognitive maps (FCMs) is proposed. It constructs a generic list of internal and external factors into FCMs to allow truster agents to infer if a trustee agent is worthy of interacting with. Each truster agent can determine the values to be given to the causal links between different factors so as to express their own preferences. Nevertheless, belief source variations and the variations in choosing values for the causal links can heavily affect the performance of the model and it is difficult to verify the validity of the models produced since there is a large degree of subjectivity involved.

The model proposed in [20] narrows down the scope of analysis to focus on the relationship between agents. The relationships used in their model are not social relationships but market relationships built up through interactions. The model identifies the relationships between agents (e.g., trade, dependency, competition, collaboration, tripartite) by analyzing their interactions through the perspective of an agent-based market model; these relationships are then filtered to identify the ones most relevant to the analysis of agent trustworthiness; then, the relationships are interpreted to derive values to estimate the trustworthiness of the agents.

The SUNNY model [33] is the first trust inference model that computes a confidence measure based on social network information. The model maps a trust network into a Bayesian Network which is useful for probabilistic reasoning. The generated Bayesian Network is then used to produce estimates of the lower and upper bounds of confidence values for trust evaluation. The confidence values are used as heuristics to calculate the most accurate estimations of the trustworthiness of the trustee agents in the Bayesian Network.

In [40], the bootstrapping problem facing evidence-based trust models is investigated. In bootstrapping, it is assumed that neither prior interaction experience nor social relationship information is available about trustee agents who are newcomers to an MAS. In this work, the underlying intuition used to design the model is that the intrinsic properties of a trustee agent can reflect its trustworthiness to some degree. The model learns a set of stereotypes based on the features in trustee agents' profiles using a decision tree based technique. Newcomer trustee agents are then classified into different stereotypes and stereotypical reputation values are produced for them. Nevertheless, due to the lack of suitable data, this paper did not point out which features may be useful in estimating a trustee agent's trustworthiness.

In [53], trust evaluation models are enriched by incorporating human dispositions such as optimism, pessimism and realism into the process of selecting whose opinions to believe in. The model proposed in this work consists of a two-layered cognitive filtering algorithm. The first layer filters out the agents who lack required experience or reliability using the BRS and the uncertainty measure proposed in [32]. The second layer calculates a similarity measure for opinions received from witness agents and the current belief by the truster agent. Combining it with the truster agent's innate disposition, the model produces credibility measures for the witness agents and enables the truster agent to know whose opinions it should trust more.

In [86], a fuzzy logic based testimony aggregation model is proposed to reduce the need for human experts to set key decision threshold values used in many heuristic reputation models. The model analyzes the temporal characteristics of the testimonies, the similarity between incoming testimonies and the current belief, and the quantity of testimonies available to determine the weight to be assigned to each testimony when computing the reputation of a trustee agent. The model was shown to be robust against Sybil attacks using data collected by [82].

4) Organizational Trust Evaluation Models

Another approach to maintaining trust in an MAS is to introduce an organizational structure into multi-agent trust management. Such a goal can be accomplished only if there exists at least one trusted third-party in an MAS who can act as a supervising body for the transactions among other agents.

In one of the earliest research works in this area [7], the proposed framework consists of three components: 1) a specific transactional organization structure made of three roles (i.e., the addressee, the counter-party and the authority), 2) a contract specification language for contract management, and 3) a set of contract templates created using the contract specification language. In order to conduct transactions, an agent needs to register with the authority, negotiate with other agents to set up the terms in the contracts, and carry out the work required by the contracts under the supervision of the authority.

The Certified Reputation (CR) model is proposed in [10]. It provides a mechanism for a trustee agent to pro-vide truster agents with certified ratings about its past performance. It is possible to make sharing certified ratings as a standard part of setting up a transaction be-tween agents. By putting the burden of demonstrating past performance on the trustee agents, truster agents can save on efforts required to solicit third-party testimonies and filtering these testimonies. In addition, the certified ratings are provided by the trustee agent's previous interaction partners, thus making the CR model a distributed approach which is suitable for use in MASs.

In [16], an agent coordination mechanism based on the interplay of trust and organizational roles for agents is proposed. It provides a mechanism for agents to establish which task a trustee agent is good at through multiple interactions and allows the role each agent can play in an agent society to gradually evolve and thus, dynamically changes the organizational structure by evolving an organizational taxonomy in the MAS. In subsequent interactions, the updated roles for the trustee agents act as references for truster agents to decide how to delegate tasks.

C. Trust-aware Interaction Decision-making

Existing trust-aware interaction decision making approaches can be broadly divided into two categories: 1) *greedy* and 2) *dynamic*. Such a classification is based on the strategy adopted by the different approaches in terms of selecting trustee agents for interaction. Greedy approaches tend to use simple rules while dynamic approaches often attempt to assess the changing conditions in the operating environment in an effort to balance the exploitation of known trustworthy trustee agents with the exploration for potentially better alternatives.

In a typical greedy approach, a truster agent explores for trustee agents with a desired reputation standing through either some supporting infrastructure (e.g., peer recommendation, social network analysis, etc.) or random exploration. The reputation values of the candidate trustee agents are calculated using a trust evaluation model of choice, and the one with the highest estimated reputation is selected for interaction. This approach is the most widely adopted in the computational trust literature [69], [9], [14], [70], [37], [71], [45], [72]. From an individual truster agent's point of view, in order to maximize its own long term wellbeing, it is advantageous to select the best available option.

Compared to static approaches, there are fewer dynamic approaches in the computational trust literature. A reinforcement learning based approach is proposed in [11]. The gain derived by a truster agent from choosing each trustee agent for interaction consists of the Q-value from Q-learning as well as the expected value of perfect information. At each time step, a truster agent chooses an action (i.e., exploration v.s. exploitation) which can maximize its gain.

In [87], the authors measure a truster agent's know-ledge degree about each potential trustee agents and use this metric to determine which trustee agent to select for interaction. The knowledge degree depends on the amount of direct past interaction experience with the trustee agent, third-party testimonies about that trustee agent, and the self reported trustworthiness by that trustee agent available to the truster agent. The value of the knowledge degree is normalized to the range [0, 1], with 1 representing ”completely known” and 0 representing ”no direct interaction experience”. In the local record of a truster agent, candidate trustee agents are organized into four different groups according to their knowledge degree values. If there are enough trustee agents with reputation values higher than a predefined threshold in the most well known group, the truster agent will only select from these trustee agents for interaction; otherwise, a number of exploration rounds will be allocated to trustee agents in groups to build up the knowledge degree about them and promote them into higher order groups.

Another dynamic approach proposed in [88] measures how much the behavior of the trustee agents has changed to determine the amount of effort a truster agent should devote to exploration. In this approach, each truster agent keeps track of the *long term trust* ($LT_i(t)$) and the *short term trust* ($ST_i(t)$) values of candidate trustee agent i, where $ST_i(t)$ reflects the changes in i's behavior faster than $LT_i(t)$. The average absolute difference between $LT_i(t)$ and $ST_i(t)$ is used to estimate the collective degree of change $C(t)$ in trustee agents' behavior. When $C(t)$ is larger than 0, an exploration extent value $E(t)$ is calculated. Together with the reputation value of each trustee agent, this value is used to derive a selection probability $RP_i(t)$ for every trustee agent. The candidate trustee agents are then selected using a Monte Carlo method based on their $RP_i(t)$ values. When $C(t) = 0$, the trustee agent with the highest reputation evaluation is always selected for interaction.

In [89], a task delegation decision approach - Global Considerations - has been proposed with the objective the reduce the delay experienced by the truster agents. Based on the effort required to effectively handle the number of incoming requests for a trustee agent i at time t ($e_{in,i}(t)$), and the effort i is able to expend at t ($e_i(t)$), the reputation of i is discounted. If $\frac{e_i(t)}{e_{in,i}(t)} \geq 1$, the probability of j being selected by a truster agent in the next iteration, $P_i(t+1)$, is directly proportional to its reputation. Otherwise, $P_i(t+1)$ is discounted by $\frac{e_i(t)}{e_{in,i}(t)}$. In this way, trustee agents whose capacities are being heavily utilized will have lower chances of being assigned more tasks in subsequent time steps.

Apart from decisions on the balance of exploration and exploitation, a truster agent can also decide on when to use additional mechanisms to induce the desired behavior from trustee agents following the framework proposed in [17]. In this framework, strategies a truster agent can adopt include 1) explicit incentives, 2) monitoring, and 3) reputational incentives. Based on the consideration of a wide range of factors including reputation, cost of monitoring, expected loss, expected value of monitoring an activity, etc., a truster agent dynamically makes a choice among these strategies in addition to the decision as to which trustee agent to select for an interaction.

While most trustworthiness evaluation models and trust-aware interaction decision-making approaches are designed for truster agents to use, [5] proposed an interesting model that includes mechanisms to help trustee agents determine how trustworthy to be. Based on the Agent Reputation and Trust (ART) testbed, the interdependencies, rewards and complexities of trust decisions are identified. A Q-learning based method is then used to help truster agents determine who to trust, how truthful to be in sharing reputation information and what reputations to believe in; and to help trustee agents to determine how trustworthy to be.

The decision to trust an agent can be considered a reasoning process as shown in [90]. In [91], a first of its kind argument scheme is proposed to facilitate trust reasoning. It takes a wide range of factors into account when reasoning about trust. These factors include direct experience, indirect experience, expert opinion, authority certification, reputation, moral nature, social standing, majority behavior, prudence, and pragmatism. Each of these factors are associated to a set of critical questions that needs to be answered in order to establish trust. The proposed scheme provides researchers with a platform to further advance reasoning-based trust-aware decision-making.

D. *Performance Assessment Methods*

To evaluate the effectiveness of a proposed model in multi-agent trust research, two major approaches are available: 1) simulation-based evaluation, and 2) evaluation through real world datasets. Each of these methods has its own merits and has been applied either individually or in combinations by researchers.

To date, the most widely used method for evaluating a trust model is through simulations. In an effort to standardize the evaluation of trust models through simulations, the Agent Reputation and Trust (ART) testbed [6] was proposed and a series of competitions using this testbed were held in the *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. The ART testbed creates an environment where agents need to delegate tasks to each other to produce appraisals for virtual artworks and earn virtual profits during this process. Nevertheless, the testbed is designed mainly for evaluating models that aim to mitigate the adverse effects of biased third-party testimonies which is only one of the problems multi-agent trust research aims to address. Three ART testbed competitions were held in AAMAS from 2006 to 2008. Currently, researchers are still creating their own simulation environments in order to produce conditions under which their trust models are designed to operate.

Another way of evaluating the performance of a trust model is based on data collected from real world applications. Depending on the specific features in a trust model, data from different types of online sources may be selected. For example, the *Epinions* dataset and the *Extended Epinions* dataset compiled by [92] have been used to analyze the performance of trust models concerned with bootstrapping and collaborative recommendation [93], [94], [46], [95]; in [33], data from the *FilmTrust* social network are used to analyze the performance of the proposed SUNNY socio-cognitive trust model; rating information from eBay is used in [38]; the web spam dataset from Yahoo! is used in [49] to evaluate their model for propagating trust and distrust in the web; and data crawled from the Internet auction site *Allegro* are used as part of the evaluation in [63].

Real world data enable researchers to have a better idea of how their models would work under realistic environment conditions. However, the behavior patterns of the users in such datasets are fixed which makes it difficult for researchers to vary experimental conditions to simulate different ways in which the model can be attacked. In addition, many datasets are not specifically collected for the purpose of evaluating trust models. Thus, they may lack the ground truth about the user behavior and intention to facilitate more in-depth analysis of the performance of proposed trust models. In order to comprehensively evaluate a trust model, we believe that a combination of the two approaches should be employed. Nevertheless, it is often difficult to collect data from real world sources or to convince the industry to release datasets related to trust research, given concerns surrounding privacy and trade secret protection. From a survey of research papers published in well-known international forums such as *the Association for the Advancement of Artificial Intelligence (AAAI) Conference*, *the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)* and *the International Joint Conference on Artificial Intelligence (IJCAI)* from 2002 to 2012, we have found that over 80% of them use simulations to assess the performance of their proposed trust models.

IV. Multi-agent Trust as a Congestion Game

Most of the existing multi-agent trust models are pro-posed for application domains where trustees represent computerized services or resources (e.g., a node in a P2P file sharing system, a video on Youtube). In these cases, the performance of a trustee agent can be consistently maintained even if it is a popular choice by the truster agents (i.e., it routinely experiences a high workload). Thus, assumption 2.h, which we refer to as the *unlimited processing capacity (UPC) assumption*, can be justified.

However, in cases where trustee agents represent human service providers (e.g., C2C e-commerce systems, crowdsourcing systems), the number of requests a human trustee agent can effectively fulfill per unit time is limited by a wide range of factors such as skills, mood, and physical condition. The UPC assumption may be unreasonable. In these cases, if a trustee is overloaded with requests, the quality of the produced results as well as the timeliness of producing these results may deteriorate, causing losses for the trusters.

For trust management in resource constrained MASs, the trust game needs to be enriched with the concept of *congestion games*. Congestion games are games in which the payoff for each player depends on the resources it selects and the number of players selecting the same resources. For example, commuting to work can be modeled as a congestion game where the time taken by an agent from its home to its workplace depends on how many other agents are taking the same public transport it chooses on each day. In a Discrete Congestion Game (DCG) [96], the following components are present:

- A base set of congestible resources $\mathbb{E}$;
- n players;
- A finite set of strategies $\mathbb{S}_i$ for each player where a strategy $P \in \mathbb{S}_i$ is a subset of $\mathbb{E}$;
- For each resource e and a vector of strategies $(P_1, ..., P_n)$, a load x_e is placed on e;
- For each resource e, a delay function d_e maps the number of players choosing e to a delay represented by a real number;
- Given a strategy P_i, player i experiences a delay of $\sum_{e \in P_i} d_e(x_e)$, assuming that each d_e is positive and monotonically increasing.

In systems with conditions similar to that described in the DCG, the performance perceived by a truster agent delegating tasks to a reputable trustee agent it has found is partially dependent on how many other truster agents are making the same choice at the same time. The key difference in our case is that the perceived performance also depends on the behavior of the trustee agent which is uncertain in an open MAS. We formulate the Congestion Trust Game (CTG) as a special type of DCG where trustee agents (resources) may behave maliciously.

[Definition 1 (*Congestion Trust Game*)]: A congestion trust game is specified by a 4-tuple $\langle \mathbb{E}, \overrightarrow{l}, \mathbb{C}(t), \mathbb{V}(t) \rangle$ where

- $\mathbb{E}$ is a finite set of trustee agents with limited task processing capacity per unit time;
- $\overrightarrow{l}$ is a vector of latency functions expressing the delay experienced by truster agents selecting the set of trustee agents for task delegation. The eth component of $\overrightarrow{l}$ is denoted as l_e. It is a non-decreasing function mapping from $\mathbb{R}^+$ to $\mathbb{R}^+$;
- $\mathbb{C}(t)$ is the set of possible connections between a finite set of truster agents $\mathbb{W}$ and trustee agents in $\mathbb{E}$ in the MAS when delegating their tasks to trustee agents in $\mathbb{E}$ at time step t. The connections depend on what types of tasks each truster agent wants to delegate to each trustee agent and the qualifications of the trustee agents to perform these tasks. For example, truster agent A may want to find a trustee agent who sells apples at time step t. A subset of trustee agents $\mathbb{C}'(t)$ are qualified for this task. At time step $(t+1)$, A is looking to buy computers. Therefore, a different subset of trustee agents $\mathbb{C}''(t+1)$ are qualified for this task. The difference between the set of possible connections between truster agents and trustee agents can also be caused by agents dynamically joining or leaving an MAS;
- $\mathbb{V}(t)$ is the set of functions for calculating the potential cost for truster agents who choose to delegate tasks to the same trustee agent at time step t.

[Definition 2 (*Task Delegation Flow*)]: A task delegation flow in the CTG is a function f mapping from $\mathbb{C}(t)$ to $\mathbb{R}^+$. It can be regarded as the amount of workload assigned to a trustee agent.

[Definition 3 (*Task Delegation Cost*)]: The load on a trustee agent $e \in \mathbb{E}$ at time step t is

$$x_e = \sum_{c(t) \in \mathbb{C}(t) | e \in c(t)} f(c(t)). \tag{3}$$

The delay on a connection $c(t) \in \mathbb{C}(t)$ is

$$L(c(t)) = \sum_{e \in c(t)} l_e(x_e(f)). \tag{4}$$

The cost of a task delegation to a truster agent is

$$v(f) = \sum_{c(t) \in \mathbb{C}(t)} \frac{f(c(t)L(c(t))}{\tau_e(t)} = \sum_{e \in \mathbb{E}} \frac{x_e(f) l_e(x_e(f))}{\tau_e(t)} \tag{5}$$

where $\tau_e(t)$ is the reputation of trustee agent e in performing a given type of task as assessed at time step t.

V. The Reputation Damage Problem

Under the framework of the original trust games, existing researches often take an individual truster agent's perspective when evaluating the effectiveness of their proposed trust models. This gives rise to a variety of individual-performance-centric evaluation metrics. Two of the most widely adopted such metrics are the *average accuracy rate* of the evaluated reputation values [15], [14], [70], [97], [28], [39], [42], [48], [45], [52], [61], [56], [62] and the *average utility* achieved by individual truster agents [98], [2], [10], [38], [26], [34], [11], [12], [50], [40], [17], [49], [54].

To achieve high individual gain, most existing trust models adopt the greedy interaction decision-making approach after evaluating the reputations of trustee agents - a truster agent selects the most trustworthy trustee agent known to it for task delegation as often as possible. This behavior is consistent with the assumption that, in an open MAS, individual agents are self-interested and do not have a common goal. In many existing trust models studied under the UPC assumption, such an approach seems to yield good performance results.

Evidence-based trust models depend on the feedback provided by truster agents in order to function. Such feedbacks are often regarded as a way to reflect the subjective belief in the quality of the result received by the truster agent from an interaction with a trustee agent. The quality of an interaction is judged by the commonly used quality-of-service (QoS) metrics suitable in the context of the interaction. It is often made up of two main factors 1) metrics related to the correctness of the interaction results, and 2) metrics related to the timeliness of delivery of the results. For example, when a truster agent A wants to send a message to another agent using the messaging service provided by trustee agent B, only if the message is successfully received by the recipient agent within the expected time will A consider the interaction with B to be successful.

Under the UPC assumption, the satisfaction of the correctness and timeliness requirements depends only on the intrinsic characteristics of the trustee agent. The collective choice of interaction partners made by a population of truster agents has no effect on the observed performance of the trustee agents. However, in congestion trust games, the timeliness aspect depends on many factors, including 1) the processing capacity (or effort level) of the trustee agent which is innate to the trustee agent, and 2) the current workload of the trustee agent which is exogenous to the trustee agent. In this situation, the trustee agent's receiving a good feedback not only depends on its own trustworthiness, but also on the collective interaction decisions made by the truster agents. Here, we assume that truster agents do not purposely distort their feedbacks.

Interaction outcome evaluation is simple with the UPC assumption. Since results can always be assumed to be received on time, a truster agent just needs to produce a rating based on the correctness of the results. Such a rating can be binary (i.e., success/failure) [69] or multi-nominal (i.e., on a scale of 1 to n) [68]. Nevertheless, existing works are generally vague on how a rating based on a received interaction result and a truster's own preferences can be derived. This is mainly

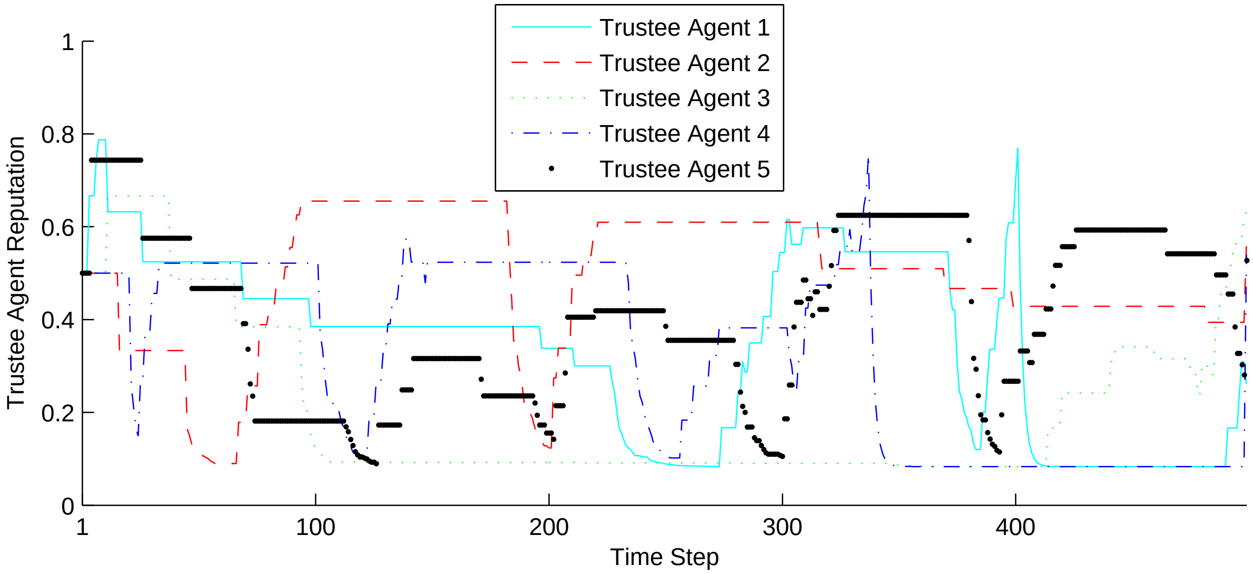


Fig. 1. Changes in reputation values of five trustee agents from the Hon group under *BRSEXT* without *clean sweep*.

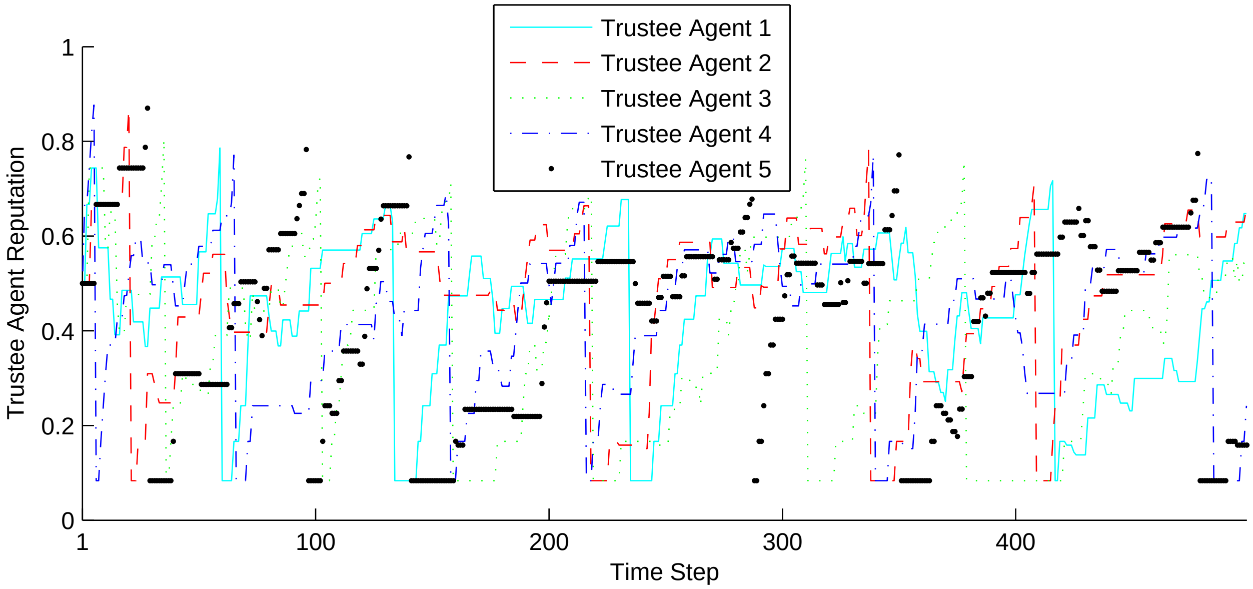


Fig. 2. Changes in reputation values of five trustee agents from the Hon group under *BRSEXT* with *clean sweep*.

because of the difficulty of designing a generic model to judge the correctness of a received result relative to an agent's preference as this may be manifested in different ways for different domains of application. For example, in an e-commerce system, receiving a parcel containing the purchased item with the expected quality may be considered a successful interaction result while, in a crowdsourcing system, receiving a file containing a piece of properly transcribed audio may be considered a successful interaction result. These ratings can be relatively easily produced by human beings, but are difficult for software agents to determine.

With the removal of the UPC assumption, the timeliness aspect of an interaction result needs to be explicitly taken into account when evaluating the outcome of that interaction. Keeping track of the deadlines of a large number of interactions is a task that is more tractable for a software agent than a human user. In this situation, the rating produced by the user based on the quality of the interaction result needs to be discounted by the timeliness of its reception in order to derive a feedback for future evaluation of the trustee's reputation.

Intuitively, if no result is received when the predetermined hard deadline has passed, the interaction should be rated as a failure. For example, agent A depends on agent B to provide it with a component in order to build a produce and deliver to agent C by a certain date T, and A needs at least N_A days to assemble the product once the component from B is received. If B fails to make the delivery by day $(T - N_A)$, there is no way for A to serve C on time. In this case, A will consider its interaction with B as a failure regardless of whether B delivers the component with high quality the following day.

In this case, the timeliness discount factor can be a binary function as:

$$f_{td}(T_{end}) = \begin{cases} 1, \text{ if } T_{end} < T_{dl} \\ 0, \text{ otherwise} \end{cases} \tag{6}$$

where T_{end} is the actual time when the interaction result is received by the truster agent, and T_{dl} is the stipulated deadline.

To further distinguish the performances of different trustees, the timeliness discount factor can be made into a smooth function with respect to the difference between T_{end} and T_{dl}. The closer T_{end} is to the time the interaction started (T_{start}), the closer $f_{td}(T_{end})$ should be to 1; the closer T_{end} is to T_{dl}, the closer $f_{td}(T_{end})$ should be to 0. A simple example of such a function may be of the form:

$$f_{td}(T_{end}) = 1 - \frac{T_{end} - T_{start}}{T_{dl} - T_{start}}. \tag{7}$$

The concept of timeliness discount can be incorporated into an existing trust model such as BRS [69] as follows:

$$\tau_{i,j}(t) = \frac{\alpha + 1}{\alpha + \beta + 2} \tag{8}$$

$$\alpha = \sum_{k=1}^{N_{i,j}} p_k, \beta = \sum_{k=1}^{N_{i,j}} n_k \tag{9}$$

$$p_k = f_{td}(T_{end}^k) \cdot O_{i \to j}(T_{end}^k) \tag{10}$$

$$n_k = f_{td}(T_{end}^k) \cdot (1 - O_{i \to j}(T_{end}^k)) + (1 - f_{td}(T_{end}^k) \tag{11}$$

where $\tau_{i,j}(t)$ is the trustworthiness evaluation of trustee agent j from the perspective of truster agent i, T_{end}^k is the completion time of the kth interaction between i and j, $N_{i,j}$ is the total number of times i delegated tasks to j, and $O_{i \to j}(t)$ is the actual outcome of i trusting j received at time t ($O_{i \to j}(t) = 1$ if the interaction is successful, otherwise, $O_{i \to j}(t) = 0$). In this paper, we use Equation (6) to calculate the timeliness discount. This extended version of BRS is referred to as *BRSEXT*.

To understand the performance of the existing trust models under the congestion trust game, an experiment is designed as follows. A simulated MAS consists of 200 trustee agents and 1,000 truster agents. This condition is similar to those found in e-commerce systems where trusters outnumber trustees by a significant margin. At each time step of the simulation, a truster agent needs to engage the services of a trustee agent in order to achieve its goal. Truster agents employ a variation of *BRSEXT* in which for a randomized 15% of the time, a truster agent will explore for potentially better alternative trustee agents by randomly selecting a trustee agent for interaction. The rest of the time, the truster agent greedily selects the known trustee agent with the highest trustworthiness value for the interaction. The trustee agent population consists of 50% of agents who produce correct results 90% (Hon) of the time on average and 50% of agents who produce correct results 10% (Mal) of the time on average. Throughout the simulation, the behavior patterns of the trustee agents do not change. A trustee agent can serve at most 10 interaction requests in its request queue per unit time. A uniform deadline of 3 time steps is used for all interaction requests. Each simulation is run for 500 time steps and the reputation values of all trustee agents are updated at each time step.

If the trustee agents are aware of the deadline requirements of the requests when the requests are accepted, they can periodically clean up their request queues to get rid of pending requests whose deadlines have passed and inform the requesting truster agents of this decision. We call this operation clean sweep. Without the clean sweep operation, the trustee agents keep work-ing on pending requests without regard to whether their deadlines have passed.

The changes in reputation values of five agents belonging to the Hon group of trustee agents without clean sweep operations are shown in Figure 1. The changes in trustee agents' reputation values as evaluated by *BRSEXT* are as follows:

1) *Reputation Building Phase*: During this phase, the agent's (for example Agent 2's) reputation starts from a low or neutral level. At this stage, not many truster agents want to interact with this Agent 2. However, due to random exploration by some truster agents, Agent 2 can get some requests. Since its reputation is relatively low compared to those of other trustee agents, the workload of Agent 2 is likely to be within a level which it can easily handle. Since Agent 2 belongs to the Hon group of trustee agents, the quality of its service is high on average. Gradually, its reputation is built up due to the positive feedbacks received from satisfied truster agents.
2) *Reputation Damage Phase*: As Agent 2 builds up its reputation, it is known to an increasing number of truster agents. More truster agents start to request its services. Gradually, the workload of Agent 2 increases past its processing capacity which results in longer delays for some requests. As more requests fail to be served within their stipulated deadlines, negative feedbacks from disgruntled truster agents start to damage Agent 2's reputation.

From Figure 1, it can be seen that the reputation values of the trustee agents alternate between these two phases. Their reputation values fluctuate around an average of 0.5272 which is 41.42% lower than their actual trustworthiness which is 0.9.

Figure 2 shows the changes in trustee agents' reputation values, as evaluated by *BRSEXT*, with the use of clean sweep operations by the trustee agents. The two phases can still be observed although the lengths of their cycles have become visibly shorter than in the case of no clean sweep operation. This is due to the fact that once a clean sweep operation is performed, the truster agents are informed of the fact that their requests cannot be served by the trustee agents. Therefore, their negative feedbacks can be issued more quickly than in the case of no clean sweep operation and have an impact on the trustee agents' reputation values. From Figure 2, it can be seen that the reputation values of the trustee agents alternate between these two phases. Their reputation values fluctuate around an average of 0.4298 which is 52.25% lower than their actual trustworthiness which is 0.9.

However, such a drastic deviation from the ground truth is not due to the fault of the trustee agents. On the contrary, they are victims of their own success. The greedy approach employed by truster agents when using the reputation evaluations to guide their interaction decisions with the aim of maximizing their own chances of success has caused the reputation evaluation to reflect not only the behavior pattern of the trustee agents, but also the impact of the collective interaction decisions made by the truster agents. Since the interaction decision-making mechanism employed by existing trust models have not taken this factor into account, this phenomenon results in instability in the MAS and negatively affects the wellbeing of trustee agents and truster agents. In this paper, we refer to this phenomenon as *Reputation Damage Problem (RDP)*.

Among the factors affecting the perceived trustworthiness of a trustee agent, the timeliness of task completion is one that is affected both by the ability and willingness of the trustee agent as well as the task delegation decisions made by the truster agents. If the RDP is not mitigated, the resulting reputation values will not fully reflect the behavior of the trustee agent, and become biased with the influence of the environment in which the trustee agent operates. In this case, the reputation value will lose its fairness and become less useful in guiding the decision making process of truster agents in subsequent interactions.

This phenomenon was first studied in [99] under crowdsourcing system conditions which resemble a congestion trust game. The study discovered that existing trust models actually result in reduction in the social welfare produced by the system, and recommended that future trust models should make an effort to distribute interaction requests fairly among reputable trustees. Their concept of distributive fairness should be distinguished from the concept of fairness proposed in [100] which states that more reputable trustees should receive more interaction opportunities than less reputable ones.

A constraint optimization based approach - SWORD - is proposed in [101] to address the RDP. It acts as a task delegation broker for truster agents to balance the workload among reputable trustee agents based on their real-time context while taking into account variations in their reputations. Nevertheless, being a centralized approach, SWORD suffers from scalability issues and a lack of decision-making transparency towards truster agents. To address this issue, a distributed constraint optimization based approach - DRAFT - is proposed in [102] to enable resource constrained trustee agents to determine which incoming task requests are to be accepted in real-time in order to protect their reputations from being damaged by the RDP. Currently, the RDP remains an open problem in need of good solutions if trust agents are to efficiently operate alongside human beings in future MASs.

VI. Discussions and Future Research

As proposed in [103] and reiterated in [104], a successful trust evaluation model should be designed with the following nine desired characteristics:

1) *Accurate for long-term performance*: The model should reflect the confidence of a given reputation value and be able to distinguish between a new entity of un-known quality and an entity with bad long-term performance.
2) *Weighted toward current behavior*: The model should recognize and reflect recent trends in entity performance.
3) *Efficient*: The model should be able to recalculate a reputation value quickly. Calculations that can be performed incrementally are important.
4) *Robust against attacks*: The model should resist attempts of any entity or entities to influence scores other than by being more honest or providing higher quality services.
5) *Amenable to statistical evaluation*: It should be easy to find outliers and other factors that can make the model produce reputation values differently.
6) *Private*: No one should be able to learn how a given witness rated an entity except the witness himself.
7) *Smooth*: Adding any single rating or a small number of ratings should not have a significant impact on the reputation value.
8) *Understandable*: The implications of the reputation value produced by a model should be easily understood by the users.
9) *Verifiable*: The recorded data should be able to show how a reputation value is calculated.

While these characteristics are helpful for designers to construct trust models that can provide useful reputation values, there is a shortage of guidelines on how interaction decisions should be made to efficiently utilize the capacities and resources the agents possess. To realize this goal, trusting decisions should achieve social equity among trustee agents in the long run. In this case, we define social equity as the situation where every trustee agent should receive interaction opportunities commensurate with its reputation. Based on the analysis in Section VI, we propose to add to the above list the following desirable characteristics for trust models in resource constrained environments:

10) *Situation-aware*: When making interaction decisions based on reputation information, the situation facing the candidate trustee agents should be taken into account so as to achieve socially equitable utilization of trustee capacities.

When human motivations interact in multi-agent systems, new research opportunities arise for trust management. Reputation rating distortions have been reported in some of the world's largest online e-commerce systems as one of the goals of people participating in such schemes is to quickly build up their reputations through illegitimate transactions [105]. Although their subsequent behaviors may not necessarily be malicious, such a practice is very disruptive to the e-commerce community. In general, reputable sellers whose reputations are gained through illegitimate means often appear to be building up their reputations much faster than peer members in the community. Integrating the analysis of the temporal aspect of the reputation building process is a research direction which has the potential to yield more effective trust evidence filtering and aggregation models. In the largest e-commerce platform in China - Taobao.com, the buyer communities are starting to respond to this problem by coming up with some rudimentary guidelines on helping buyers spot collusive sellers through looking at their historical reputation scores. Nevertheless, these collusive sellers are adapting to these self-protection mechanisms by slowing down the rate at which their reputation scores grow through less greedy behavior. Future research attempts in incorporating temporal analysis into trust evidence filtering and aggregation models should therefore, be rooted in analyzing real world data to identify and respond to these complex behavior patterns.

As human beings with limited resources (in terms of task processing capacity, time, health, etc.) are starting to play the role of trustees in many online communities (e.g., e-commerce systems, crowdsourcing systems), research on modeling their utility functions using a human centric approach is necessary for trust-aware interaction decision-making mechanisms. For example, a decision model that takes the overall wellbeing of a human trustee into account may not necessarily adopt a utility function which is always linearly related to the amount of work allocated to him. Instead, a more complex utility function that allows the decision model to vary the delegation of tasks to trustees in such a way as to achieve work/life balance for the trustees, while satisfying the overall goal of the community, will be desirable in future human-agent collectives.

Acknowledgment

This research is supported in part by Interactive and Digital Media Programme Office (IDMPO), National Research Foundation (NRF) hosted at Media Development Authority (MDA) of Singapore under Grant No.: MDA/IDM/2012/8/8-2 VOL 01.

References

[1] H. Yu, Z. Shen, C. Miao, C. Leung, and D. Niyato, "A survey of trust and reputation management systems in wireless communications," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1755–1772, 2010.

[2] T. Tran and R. Cohen, "Improving user satisfaction in agent-based electronic marketplaces by reputation modelling and adjustable product quality," in *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04)*, vol. 2, 2004, pp. 828–835.

[3] T. Tran, "A reputation-oriented reinforcement learning approach for agents in electronic marketplaces," in *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-02)*, 2002, pp. 989–989.

[4] C. Castelfranchi, R. Falcone, and G. Pezzulo, "Trust in information sources as a source for trust: a fuzzy approach," in *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'03)*, 2003, pp. 89–96.

[5] K. K. Fullam and K. S. Barber, "Learning trust strategies in reputation exchange networks," in *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'06)*, 2006, pp. 1241–1248.

[6] K. K. Fullam, T. B. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. S. Rosenschein, L. Vercouter, and M. Voss, "A specification of the agent reputation and trust (art) testbed: experimentation and competition for trust in agent societies," in *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'05)*, 2005, pp. 512–518.

[7] M. J. Kollingbaum and T. J. Norman, "Supervised interaction: Creating a web of trust for contracting agents in electronic environments," in *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'02)*, vol. 1, 2002, pp. 272–279.

[8] N. Griffiths, "Task delegation using experience-based multi-dimensional trust," in *Proceedings of the 4th international joint conference on Autonomous agents and multiagent systems (AAMAS'05)*, 2005, pp. 489–496.

[9] B. Yu and M. P. Singh, "Detecting deception in reputation management," in *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'03)*, 2003, pp. 73–80.

[10] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt, "Certified reputation: How an agent can trust a stranger," in *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'06)*, 2006, pp. 1217–1224.

[11] W. T. L. Teacy, G. Chalkiadakis, A. Rogers, and N. R. Jennings, "Sequential decision making with untrustworthy service providers," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'08)*, vol. 2, 2008, pp. 755–762.

[12] R. Kerr and R. Cohen, "Smart cheaters do prosper: Defeating trust and reputation systems," in *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'09)*, vol. 2, 2009, pp. 993–1000.

[13] R. Falcone and C. Castelfranchi, "Trust dynamics: How trust is influenced by direct experiences and by trust itself," in *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04)*, vol. 2, 2004, pp. 740–747.

[14] W. T. L. Teacy, J. Patel, N. R. Jennings, and M. Luck, "Coping with inaccurate reputation sources: Experimental analysis of a probabilistic trust model," in *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'05)*, 2005, pp. 997–1004.

[15] R. Falcone, G. Pezzulo, C. Castelfranchi, and G. Calvi, "Why a cognitive trustier performs better: Simulating trust-based contract nets," in *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04)*, vol. 3, 2004, pp. 1394–1395.

[16] R. Hermoso, H. Billhardt, and S. Ossowski, "Role evolution in open multi-agent systems as an information source for trust," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, vol. 1, 2010, pp. 217–224.

[17] C. Burnett, T. J. Norman, and K. Sycara, "Trust decision-making in multi-agent systems," in *Proceedings of the 22th International Joint Conference on Artifical Intelligence (IJCAI'11)*, 2011, pp. 115–120.

[18] P. Bedi, H. Kaur, and S. Marwaha, "Trust based recommender system for the semantic web," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, 2007, pp. 2677–2682.

[19] K. Regan, P. Poupart, and R. Cohen, "Bayesian reputation modeling in e-marketplaces sensitive to subjecthity, deception and change," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, vol. 2, 2006, pp. 1206–1212.

[20] R. Ashri, S. D. Ramchurn, J. Sabater, M. Luck, and N. R. Jennings, "Trust evaluation through relationship analysis," in *Proceedings of the 4th international joint conference on Autonomous agents and multiagent systems (AAMAS'05)*, 2005, pp. 1005–1011.

[21] J. Pasternack and D. Roth, "Making better informed trust decisions with generalized fact-finding," in *Proceedings of the 22th International Joint Conference on Artifical Intelligence (IJCAI'11)*, 2011, pp. 2324–2329.

[22] Y. Wang and M. P. Singh, "Trust representation and aggregation in a distributed agent system," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, vol. 2, 2006, pp. 1425–1430.

[23] P. Dondio and S. Barrett, "Presumptive selection of trust evidence," in *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'07)*, 2007.

[24] S. Casare and J. Sichman, "Towards a functional ontology of reputation," in *Proceedings of the 4th international joint conference on Autonomous agents and multiagent systems (AAMAS'05)*, 2005, pp. 505–511.

[25] N. Osman and D. Robertson, "Dynamic verification of trust in distributed open systems," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, 2007, pp. 1440–1445.

[26] K. K. Fullam and K. S. Barber, "Dynamically learning sources of trust information: Experience vs. reputation," in *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'07)*, 2007, pp. 1055–1060.

[27] P. Massa and P. Avesani, "Controversial users demand local trust metrics: an experimental study on epinions.com community," in *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, vol. 1, 2005, pp. 121–126.

[28] S. Reece, S. Roberts, A. Rogers, and N. R. Jennings, "A multi-dimensional trust model for heterogeneous contract observations," in

Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07), vol. 1, 2007, pp. 128–135.

[29] P. Hendrix and B. J. Grosz, "Reputation in the venture games," in *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07)*, vol. 2, 2007, pp. 1866–1867.

[30] Y. Katz and J. Golbeck, "Social network-based trust in prioritized default logic," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, vol. 2, 2006, pp. 1345–1350.

[31] T. Kawamura, S. Nagano, M. Inaba, and Y. Mizoguchi, "Mobile service for reputation extraction from weblogs: Public experiment and evaluation," in *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07)*, vol. 2, 2007, pp. 1365–1370.

[32] Y. Wang and M. P. Singh, "Formal trust model for multiagent systems," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, 2007, pp. 1551–1556.

[33] U. Kuter and J. Golbeck, "Using probabilistic confidence models for trust inference in web-based social networks," in *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07)*, vol. 2, 2007, pp. 1377–1382.

[34] S. Reches, P. Hendrix, S. Kraus, and B. J. Grosz, "Efficiently determining the appropriate mix of personal interaction and reputation information in partner choice," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'08)*, vol. 2, 2008, pp. 583–590.

[35] A. D. Procaccia, Y. Bachrach, and J. S. Rosenschein, "Gossip-based aggregation of trust in decentralized reputation systems," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, 2007, pp. 1470–1475.

[36] J. O'Donovan, B. Smyth, V. Evrim, and D. McLeod, "Extracting and visualizing trust relationships from online auction feedback comments," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, 2007, pp. 2826–2831.

[37] W. T. L. Teacy, N. R. Jennings, A. Rogers, and M. Luck, "A hierarchical bayesian trust model based on reputation and group behaviour," in *the 6th European Workshop on Multi-Agent Systems*, 2008.

[38] A. Rettinger, M. Nickles, and V. Tresp, "A statistical relational model for trust learning," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'08)*, vol. 2, 2008, pp. 763–770.

[39] C.-W. Hang, Y. Wang, and M. P. Singh, "Operators for propagating trust and their evaluation in social networks," in *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'09)*, vol. 2, 2009, pp. 1025–1032.

[40] C. Burnett, T. J. Norman, and K. Sycara, "Bootstrapping tust evaluations through stereotypes," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, vol. 1, 2010, pp. 241–248.

[41] B. Khosravifar, M. Gomrokchi, J. Bentahar, and P. Thiran, "Maintenance-based trust for multi-agent systems," in *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'09)*, vol. 2, 2009, pp. 1017–1024.

[42] P.-A. Matt, M. Morge, and F. Toni, "Combining statistics and arguments to compute trust," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, vol. 1, 2010, pp. 209–216.

[43] J. Tang, S. Seuken, and D. C. Parkes, "Hybrid transitive trust mechanisms," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, vol. 1, 2010, pp. 233–240.

[44] A. Koster, J. Sabater-Mir, and M. Schorlemmer, "Talking about trust in heterogeneous multi-agent systems," in *Proceedings of the 22th International Joint Conference on Artifical Intelligence (IJCAI'11)*, 2011, pp. 2820–2821.

[45] S. Liu, J. Zhang, C. Miao, Y.-L. Theng, and A. C. Kot, "iclub: An integrated clustering-based approach to improve the robustness of reputation systems," in *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'11)*, 2011, pp. 1151–1152.

[46] L. Li and Y. Wang, "Subjective trust inference in composite services," in *Proceedings of the 24th National Conference on Artificial Intelligence (AAAI-10)*, 2010, pp. 1377–1384.

[47] A. Salehi-Abari and T. White, "Trust models and con-man agents: From mathematical to empirical analysis," in *Proceedings of the 24th National Conference on Artificial Intelligence (AAAI-10)*, 2010, pp. 842–847.

[48] G. Vogiatzis, I. MacGillivray, and M. Chli, "A probabilistic model for trust and reputation," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, vol. 1, 2010, pp. 225–232.

[49] X. Zhang, Y. Wang, N. Mou, and W. Liang, "Propagating both trust and distrust with target differentiation for combating web spam," in *Proceedings of the 25th National Conference on Artificial Intelligence (AAAI-11)*, 2011, pp. 1292–1297.

[50] G. Liu, Y. Wang, and M. A. Orgun, "Optimal social trust path selection in complex social networks," in *Proceedings of the 24th National Conference on Artificial Intelligence (AAAI-10)*, 2010, pp. 1391–1398.

[51] H. Fang, J. Zhang, M. Sensoy, and N. M. Thalmann, "Sarc: Subjectivity alignment for reputation computation," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12)*, vol. 3, 2012, pp. 1365–1366.

[52] X. Liu and A. Datta, "A trust prediction approach capturing agents' dynamic behavior," in *Proceedings of the 22th International Joint Conference on Artifical Intelligence (IJCAI'11)*, 2011, pp. 2147–2152.

[53] Z. Noorian, S. Marsh, and M. Fleming, "Multi-layer cognitive filtering by behavioral modeling," in *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'11)*, vol. 2, 2011, pp. 871–878.

[54] Y. Haghpanah and M. desJardins, "Prep: a probabilistic reputation model for biased societies," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12)*, vol. 1, 2012, pp. 315–322.

[55] J. Witkowski, "Trust mechanisms for online systems," in *Proceedings of the 22th International Joint Conference on Artifical Intelligence (IJCAI'11)*, 2011, pp. 2866–2867.

[56] A. Koster, J. Sabater-Mir, and M. Schorlemmer, "Personalizing communication about trust," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12)*, vol. 1, 2012, pp. 517–524.

[57] M. P. Singh, "Trust as dependence: a logical approach," in *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'11)*, vol. 2, 2011, pp. 863–870.

[58] S. Liu, A. C. Kot, C. Miao, and Y.-L. Theng, "A dempster-shafer theory based witness trustworthiness model," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12)*, vol. 3, 2012, pp. 1361–1362.

[59] M. Venanzi, M. Piunti, R. Falcone, and C. Castelfranchi, "Facing openness with socio-cognitive trust and categories," in *Proceedings of the 22th International Joint Conference on Artifical Intelligence (IJCAI'11)*, 2011, pp. 400–405.

[60] C. Burnett and N. Oren, "Sub-delegation and trust," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12)*, vol. 3, 2012, pp. 1359–1360.

[61] S. Jiang, J. Zhang, and Y. S. Ong, "A multiagent evolutionary framework based on trust for multiobjective optimization," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12)*, vol. 1, 2012, pp. 299–306.

[62] M. Piunti, M. Venanzi, R. Falcone, and C. Castelfranchi, "Multimodal trust formation with uninformed cognitive maps (uncm)," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12)*, vol. 3, 2012, pp. 1241–1242.

[63] X. Liu and A. Datta, "Modeling context aware dynamic trust using hidden markov model," in *Proceedings of the 26th National Conference on Artificial Intelligence (AAAI-12)*, 2012, pp. 1938–1944.

[64] E. Serrano, M. Rovatsos, and J. Botia, "A qualitative reputation system for multiagent systems with protocol-based communication," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'12)*, vol. 1, 2012, pp. 307–314.

[65] S. P. Marsh, "Formalizing trust as a computational concept," Ph.D. dissertation, 1994.

[66] A. J. Jones and J. Pitt, "On the classification of emotions, and its relevance to the understanding of trust," in *Workshop on Trust in Agent Societies (TRUST-11)*, 2011.

[67] D. G. Mikulski, F. L. Lewis, E. Y. Gu, and G. R. Hudas, "Trust dynamics in multi-agent coalition formation," in *SPIE - Unmanned Systems Technology*, vol. 8045, 2011.

[68] A. Jøang and J. Haller, "Dirichlet reputation systems," in *2nd International Conference on Availability, Reliability and Security (ARES)*, 2007, pp. 112–119.

[69] A. Jøang and R. Ismail, "The beta reputation system," in *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002, pp. 41–55.

[70] J. Weng, C. Miao, and A. Goh, "An entropy-based approach to protecting rating systems from unfair testimonies," *IEICE - Transactions on Information and Systems*, vol. E89-D, no. 9, pp. 2502–2511, 2006.

[71] J. Weng, Z. Shen, C. Miao, A. Goh, and C. Leung, "Credibility: How agents can handle unfair third-party testimonies in computational trust models," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 22, no. 9, pp. 1286–1298, 2010.

[72] Z. Shen, H. Yu, C. Miao, and J. Weng, "Trust-based web-service selection in virtual communities," *Journal for Web Intelligence and Agent Systems (WIAS)*, vol. 9, no. 3, pp. 227–238, 2011.

[73] H. Yu, S. Liu, A. C. Kot, C. Miao, and C. Leung, "Dynamic witness selection for trustworthy distributed cooperative sensing in cognitive radio networks," in *Proceedings of the 13th IEEE International Conference on Communication Technology (ICCT)*, 2011, pp. 1–6.

[74] C. M. Jonker and J. Treur, "Formal analysis of models for the dynamics of trust based on experiences," in *9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World: MultiAgent System Engineering (MAAMAW '99)*, 1999, pp. 221–231.

[75] M. Schillo, P. Funk, I. Stadtwald, and M. Rovatsos, "Using trust for detecting deceitful agents in artificial societies," *Journal of Applied Artificial Intelligence*, vol. 14, no. 8, pp. 825–848, 2000.

[76] J. Shi, G. V. Bochmann, and C. Adams, "Dealing with recommendations in a statistical trust model," in *Workshop on Trust in Agent Societies in conjunction with the 4th International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS'05)*, 2005, pp. 144–155.

[77] K. S. Barber and J. Kim, *Soft Security: Isolating Unreliable Agents from Society*, 2003, vol. 2631, pp. 224–233.

[78] L. Mui and M. Mohtashemi, "A computational model of trust and reputation," in *35th Annual Hawaii International Conference on System Sciences (HICSS'02)*, vol. 7, 2002, pp. 188–197.

[79] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[80] J. Sabater and C. Sierra, "Reputation and social network analysis in multi-agent systems," in *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'02)*, 2002, pp. 475–482.

[81] A. Whitby, A. Jøang, and J. Indulska, "Filtering out unfair ratings in bayesian reputation systems," pp. 1260–1262, 2004.

[82] Y. Liu and Y. Sun, "Anomaly detection in feedback-based reputation systems through temporal and correlation analysis," in *Proceedings of the 2010 IEEE Second International Conference on Social Computing (SOCIALCOM'10)*, 2010, pp. 65–72.

[83] E. S. Page, "Continuous inspection scheme," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[84] T. Qin, H. Yu, C. Leung, Z. Shen, and C. Miao, "Towards a trust aware cognitive radio architecture," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 13, no. 2, pp. 86–95, 2009.

[85] Y. Wang, C.-W. Hang, and M. P. Singh, "A probabilistic approach for maintaining trust based on evidence," *Journal of Artificial Intelligence Research (JAIR)*, vol. 40, no. 1, pp. 221–267, 2011.

[86] S. Liu, H. Yu, C. Miao, and A. C. Kot, "A fuzzy logic based reputation model against unfair ratings," in *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'13)*, 2013.

[87] V. Muñoz, J. Murillo, B. López, and D. Busquets, "Strategies for exploiting trust models in competitive multi-agent systems," in *Proceedings of the 7th German Conference on Multiagent System Technologies (MATES'09)*, 2009, pp. 79–90.

[88] M. Hoogendoorn, S. W. Jaffry, and J. Treur, "Exploration and exploitation in adaptive trust-based decision making in dynamic environments," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (IAT'10)*, 2010, pp. 256–260.

[89] A. Grubshtein, N. Gal-Oz, T. Grinshpoun, A. Meisels, and R. Zivan, "Manipulating recommendation lists by global considerations," in *Proceedings of the 2nd International Conference on Agents and Artificial Intelligence (ICAART'10)*, 2010, pp. 135–142.

[90] C. Castelfranchi and R. Falcone, "Trust is much more than subjective probability: Mental components and sources of trust," in *Proceedings of the 33rd Hawaii International Conference on System Sciences - Volume 6*, ser. HICSS'00, 2000, pp. 6008–6021.

[91] S. Parsons, K. Atkinson, K. Haigh, K. Levitt, P. McBurney, J. Rowe, M. P. Singh, and E. I. Sklar, "Argument schemes for reasoning about trust," in *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA)*, 2012, p. 430441.

[92] P. Massa and P. Avesani, "Trust-aware bootstrapping of recommender systems," in *Proceedings of the ECAI 2006 Workshop on Recommender Systems*, 2006, pp. 29–33.

[93] ——, "Trust metrics on controversial users: Balancing between tyranny of the majority and echo chambers," *International Journal on Semantic Web and Information Systems*, vol. 3, no. 1, pp. 1–21, 2007.

[94] ——, *Trust Metrics in Recommender Systems*. Springler, 2009, pp. 259–285.

[95] S. Ray and A. Mahanti, "Improving prediction accuracy in trust-aware recommender systems," in *43rd Hawaii International Conference on System Sciences*, 2010, pp. 1–9.

[96] D. Monderer and L. S. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, pp. 124–143, 1996.

[97] J. Weng, C. Miao, A. Goh, Z. Shen, and R. Gay, "Trust-based agent community for collaborative recommendation," in *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'06)*, 2006, pp. 1260–1262.

[98] R. K. Dash, S. D. Ramchurn, and N. R. Jennings, "Trust-based mechanism design," in *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04)*, vol. 2, 2004, pp. 748–755.

[99] H. Yu, Z. Shen, C. Miao, and B. An, "Challenges and opportunities for trust management in crowdsourcing," in *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'12)*, 2012, pp. 486–493.

[100] A. Wierzbicki and R. Nielek, "Fairness emergence in reputation systems," *Journal of Artificial Societies and Social Simulation*, vol. 14, no. 1, 2011.

[101] H. Yu, Z. Shen, C. Miao, and B. An, "A reputation-aware decision-making approach for improving the efficiency of crowdsourcing systems," in *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'13)*, 2013.

[102] H. Yu, C. Miao, B. An, C. Leung, and V. R. Lesser, "A reputation management model for resource constrained trustee agents," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13)*, 2013.

[103] R. Dingledine, M. J. Freedman, and D. Molnar, *Accountability*. O'Reilly Publishers, 2000.

[104] A. Jøang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision Support Systems*, vol. 43, no. 2, pp. 618–644, 2007.

[105] Hexun.com, "Sellers faking reputation ratings on taobao.com agitate buyers (translated title)," *http://tech.hexun.com/2012-08-14/144719859.html*, 2012.

Han Yu is currently a Ph.D. student in the School of Computer Engineering, Nanyang Technological University (NTU), Singapore. He is a Singapore Millennium Foundation (SMF) Ph.D. scholar. He obtained his B.Eng. in Computer Engineering with 1st Class Honours from NTU in 2007. From 2007 to 2008, he worked as a systems engineer in Hewlett-Packard Singapore Pte Ltd. In 2011, he won the Best Paper Award in the 13th IEEE International Conference on Communication Technologies (ICCT). His research interests include trust management in multi-agent systems and intelligent agent augmented interactive digital media in education.

Zhiqi Shen is currently with the School of Computer Engineering, Nanyang Technological University, Singapore. He obtained B.Sc. in Computer Science and Technology from Peking University, M.Eng. in Computer Engieering in Beijing University of Technology, and PhD in Nanyang Technological University respectively.

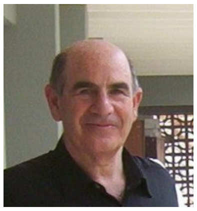
His research interests include Artificial Intelligence, Software Agents, Multi-agent Systems (MAS); Goal Oriented Modeling, Agent Oriented Software Engineering; Semantic Web/Grid, e-Learning, Bio-informatics and Bio-manufacturing; Agent Augmented Interactive Media, Game Design, and Interactive Storytelling.

Cyril Leung is a member of the IEEE and the IEEE Computer Society. He received the B.Sc.(honours) degree from Imperial College, University of London, England, in 1973, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University in 1974 and 1976 respectively.

From 1976 to 1979 he was an Assistant Professor in the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. During 1979-1980 he was with the Department of Systems Engineering and Computing Science, Carleton University, Ottawa, Canada. Since July 1980, he has been with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, B.C., Canada, where he is a Professor and currently holds the PMC-Sierra Professorship in Networking and Communications. He is the deputy director of the NTU-UBC Joint Research Centre of Excellence in Active Living for the Elderly (LILY). His current research interests are in wireless communications systems. He is a member of the Association of Professional Engineers and Geoscientists of British Columbia, Canada.

Chunyan Miao is an Associate Professor in the School of Computer Engineering (SCE) at Nanyang Technological University (NTU). She is the director of the NTU-UBC Joint Research Centre of Excellence in Active Living for the Elderly (LILY). Prior to joining NTU, she was an Instructor and Postdoctoral Fellow at the School of Computing, Simon Fraser University, Canada.

Her major research focus is on studying the cognitive and social characteristics of intelligent agents in multi-agent and distributed AI/CI systems, such as trust, emotions, motivated learning, ecological and organizational behavior. She has made significant contributions in the integration of the above research into emerging technologies such as interactive digital media (e.g., virtual world, social networks, and massively multi-player game), cloud computing, mobile communication, and humanoid robots.



Victor R. Lesser received his B.A. in Mathematics from Cornell University in 1966, and the Ph.D. degree in Computer Science from Stanford University in 1973. He then was a post-doc/research scientist at Carnegie-Mellon University, working on the Hearsay-II speech understanding system. He has been a professor in the School of Computer Science at the University of Massachusetts Amherst since 1977, and was named Distinguished University Professor of Computer Science in 2009. His major research focus is on the control and organization of complex AI systems. Professor Lesser is a Founding Fellow of the American Association of Artificial Intelligence (AAAI), and is considered a leading researcher in the areas of blackboard systems, multi-agent/ distributed AI, and real-time AI. He has also made contributions in the areas of computer architecture, signal understanding, diagnostics, plan recognition, and computer-supported cooperative work. He has worked in application areas such as sensor networks for vehicle tracking and weather monitoring, speech and sound understanding, information gathering on the internet, peer-to-peer information retrieval, intelligent user interfaces, distributed task allocation and scheduling, and virtual agent enterprizes.