Trust, Influence and Reputation Management Based on Human Reasoning

Mehrdad Nojoumian
Department of Computer Science
Southern Illinois University, Carbondale, Illinois 62901, USA
nojoumian@cs.siu.edu

Abstract

Understanding trust, influence and reputation and constructing computational models of these notions are two essential scientific challenges in computer science as well as social sciences. Although scientists in both disciplines have independently conducted research on these topics over the last couple of decades, there is a huge gap between two literatures. This paper therefore illustrates an interdisciplinary work-in-progress on trust, influence and reputation modeling based on human reasoning. Using a survey-based data collection approach, we would like to understand how humans gain/lose trust in their daily life interactions and how behavior/attitudes of humans can be influenced or shaped in various social encounters. The data will be then transformed into mathematical models to be used in technological or software systems.

Introduction

From a social science perspective, *trust* is the willingness of a person to become vulnerable to the actions of another person irrespective of the ability to control those actions (Mayer, Davis, and Schoorman 1995) and *influence* refers to any tactic used to alter the behavior or attitude of other people. However, in the computer science community, *trust* is defined as a personal expectation that a player has with respect to the future behavior of another party, i.e., a personal quantity measured to help the players in their future dyadic encounters. On the other hand, *reputation* is the perception that players have with respect to another player's intention, i.e., a social quantity computed based on the actions of a given player and the observations made by other parties in an e-community (Mui, Mohtashemi, and Halberstadt 2002). From another perspective (Castelfranchi and Falcone 1998), *trust* is made up of underlying beliefs and it is a function based on the values of these beliefs. Similarly, *reputation* is a social notion of trust. Note that trust can be formed based on local or social evidence. In the former case, trust is built by direct observations whereas, in the latter case, it is built through information from other parties, a.k.a *referral chain*.

In other words, the goal of *reputation systems* is to collect, distribute and aggregate feedback about participants' past behavior. They seek to address the development of reputation by recording the behavior of the parties. For instance, in e-commerce, the model of reputation is constructed from a buying agent's positive/negative past experiences with the aim of predicting how satisfied a buying agent will be in the future interactions with a selling agent (Resnick et al. 2000).

Our Motivation and Objectives

Trust, influence and reputation are three fundamental human factors that have been widely used in technological systems. In the social science community, there exist many fascinating discoveries and hypotheses about these notions. On the other hand, in the computer science discipline, there are many computational models of these concepts, however, these models are mainly context-oriented and they have been designed and improved based on engineering methods.

We are therefore motivated to bridge the gap between these two literatures. We would like to construct new computational models of trust, influence and reputation while incorporating human reasoning factors into the specifications of our models. We intend to utilize our constructed models in different contexts such as cryptographic constructions, as we did in the past (Nojoumian and Stinson 2012a; 2012b; Nojoumian, Stinson, and Grainger 2010), e-commerce, etc. For further motivation, the following scenarios are provided:

- Cybersecurity: consider a cyberwar setting where attackers try to launch an attack on critical infrastructures and defenders try to protect them. Here we are dealing with humans and a model that emulates human reasoning would be an appropriate choice for trust management.
- E-commerce: consider a recommender system in an online shopping setting where the reputation scores are computed based on the assigned ratings. People from different cultures and regions have different interpretations when it comes to rating and gaining/losing trust. Therefore, an appropriate model can be used to address these issues.
- Robotics: consider a group of humanoid robots that are interacting in order to accomplish certain tasks. Different models of trust and influence can be utilized based on the culture and region that each group of agents represents.

Our approach incorporates certain human reasoning factors into the specification of models when it's required. As a result, models will be adaptive to culture, region, gender, etc.

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Therefore, our high-level objectives are as follows: To bridge the gap between two different literatures on trust and influence in computer and social sciences. To investigate how trust measurement and influence work in humans. To perform cross-cultural data collection targeting samples from countries in the East and West. To quantify our data through hypothesizing and modeling. Finally, to deploy our models in technological or software systems.

Human Reasoning and Data Collection

Appropriate data collection mechanisms are required to understand how humans gain/lose trust or how their behaviors might be changed in their social encounters. One way of designing such a mechanism is to prepare a set of narratives (similar to events in our daily lives) each of which followed by questions. This is what we did in the preliminary phase. We will further illustrate our data collection methodology.

Informal Cross-Cultural Observations

We are interested in cross-cultural data analysis due to our initial observations, as explained here, and current dyadic encounters between the East and West. These observations are related to two distinct regions. They show how trust and influence might be affected by history, geopolitical factors, geography, weather, population, wealth of the area, etc. Note that these observations are made by informal questionnaires and social interactions with people in these two regions.

People in the first region were always in "caution-mode" in the first dyadic encounter; initial trust was not positive in the first few interactions; trust reduction had a sharper slope compared to trust escalation; rebuilding trust was extremely tough and took a long time; and changing people's attitude was hard when it came to influence. People in the second region were always in "question-mode" in the first dyadic encounter; initial trust value was positive in the first few interactions; trust reduction and escalation had a similar slope; rebuilding trust was not that difficult and sometimes easy; and changing people's attitude wasn't hard.

Our preliminary analysis indicated that the first region has gone through devastating drought and destructive wars in the last centuries. The financial situation has also been unstable for years. Furthermore, people have always had concerns for water, food and resources. On the other hand, the second region has been relatively safe with numerous resources, however, due to the low population density, social interactions are very limited in this area. Although these justifications might be reasonable, we intend to scrutinize them further.

Data Collection Methodology

Since we intend to analyze trust, influence and reputation from intelligence and behavior perspectives to construct computational models that emulate human reasoning, our scenarios and questionnaires must have a specific structure and capture certain relevant information. As shown below, our questions are categorized in eight distinct groups by considering cognitive, linguistic and psychological factors.

Note that demographic data will be collected at the beginning of the data collection to obtain a profile of our human subjects in terms of mindset, appearance, self-confidence, past experiences when it comes to trust/mistrust, and so on.

1. **Initial Trust:** scenarios that illustrate the initial trust values in the first dyadic encounters, e.g., when two people meet for the first time or after online chatting, or when a common friend introduces two people to each other, etc.
2. **Trust Escalation:** scenarios that illustrate a sequence of incidents/actions that escalate trust between two parties, e.g., helping, supporting, lending money, etc.
3. **Trust Reduction:** scenarios that illustrate a sequence of incidents/actions that damage trust between two parties, e.g., lying, lying for the second time, cheating, etc.
4. **Trust Mutation:** second and third groups can be a sequence of mild incidents (e.g., lying) followed by critical incidents (e.g., cheating) and vice versa.
5. **Re-Building Trust:** scenarios that demonstrate how trust can be re-built between two parties, e.g., lying, ignoring and then apologizing, paying attention, supporting, etc.
6. **Gaining Influence:** scenarios that illustrate how people might be able to change the attitude of others and the ratio of this impact over time, e.g., logically convincing someone, helping, lending money, providing useful advice, etc.
7. **Losing Influence:** scenarios that illustrate how people may lose their influential impacts on others and the ratio of this failure over time, e.g., forcing, misleading, etc.
8. **Influence Mutation:** sixth and seventh groups can be a sequence of mild incidents (e.g., misleading) followed by critical incidents (e.g., forcing) and vice versa.

For further clarification, consider the following sample questionnaire. Let trust be a value in $[-1, +1]$ (-1 and $+1$ mean fully untrustworthy and fully trustworthy respectively) and let the initial trust value for a newcomer be 0, i.e., neutral. We consider nine options for each question: fully untrustworthy (U), high negative (H^-), medium negative (M^-), low negative (L^-), neutral (N), low positive (L^+), medium positive (M^+), high positive (H^+), fully trustworthy (T).

Sample Scenario: *Alice and Bob have been in a long-term relationship and fully trust each other, i.e., the initial trust value is* $+1$. *If you were Alice, what level of trustworthiness do you assign to Bob after each of the following incidents?*

1. *Initially, Alice finds out Bob has lied to her about his salary.*

   ○U ○H^- ○M^- ○L^- ○N ○L^+ ○M^+ ○H^+ ○T
2. *Afterwards, Alice finds out Bob has lied to her about his family.*

   ○U ○H^- ○M^- ○L^- ○N ○L^+ ○M^+ ○H^+ ○T
3. *In the third incident, Ted tells Alice that Bob has cheated on her.*

   ○U ○H^- ○M^- ○L^- ○N ○L^+ ○M^+ ○H^+ ○T
4. *In the last incident, Alice herself sees Bob is cheating on her.*

   ○U ○H^- ○M^- ○L^- ○N ○L^+ ○M^+ ○H^+ ○T

By collaboration with social scientists, we were able to define a list of positive and/or negative keywords associated with trust and influence, among which are: + lending, + helping, + supporting, + giving, - lying, - cheating, - ignoring, - misleading, +/- judging, +/- relying, +/- suggesting, etc.

Hypothesizing and Modeling

We now explain four systematic steps of trust modeling that we have used in the past (Nojoumian and Lethbridge 2008; Nojoumian 2012). We plan to use this method in the future.

A. Specification

The following model specification was inspired by our preliminary/pilot data collection. We used scenarios similar to the one presented in the previous section. All actions are categorized into two groups: *cooperative* and *defective*. Parties are also classified in three groups (bad players, newcomers and good players) based on their current trust values.

(A) If a bad party cooperates, he is encouraged by a small reward, e.g., $x_e \in (0.01, 0.05)$. (B) If a newcomer cooperates, he is rewarded, e.g., by $x_g = 0.05$. (C) If a good party cooperates, he is rewarded by a factor more than the encouragement factor x_e, e.g., $x_r \in (0.05, 0.09)$. (D) If a good party defects, he is discouraged by a small penalty, e.g., $x_d \in (-0.05, -0.01)$. (E) If a newcomer defects, he is penalized, e.g., by $x_t = -0.05$. (F) If a bad party defects, he is penalized by a factor more than the discouragement factor x_d, e.g., $x_p \in (-0.09, -0.05)$.

As stated, if good and bad players cooperate, good players are rewarded more than bad players, or if good and bad players defect, good players are penalized less than bad players, etc. This is similar to the way that humans gain or lose trust in their social interactions. This model outperforms many existing schemes since it creates a reasonable trust margin between cooperative and non-cooperative parties.

B. Transformation

It is easy to transform the above specification into a mathematical model. Suppose $P_i \in \mathcal{B}$ if $\mathcal{T}_i(p) \in [-1, \beta)$, $P_i \in \mathcal{N}$ if $\mathcal{T}_i(p) \in [\beta, \alpha]$ and $P_i \in \mathcal{G}$ if $\mathcal{T}_i(p) \in (\alpha, +1]$. Let $\ell_i = 1$ denotes P_i has cooperated and $\ell_i = 0$ denotes he has defected. The proposed trust function, as shown in Figure 1, is as follows, where $x = \mathcal{T}_i(p-1)$:

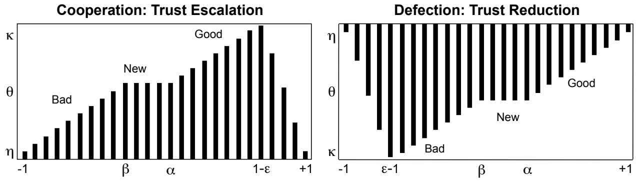

Figure 1: Trust Adjustment by $\mu(x)$ and $\mu'(x)$ Functions

$$\ell_i = 1 \quad \Rightarrow \quad \mathcal{T}_i(p) = \mathcal{T}_i(p-1) + \mu(x)$$

$$\mu(x) = \begin{cases} \dfrac{\theta - \eta}{\beta + 1}(x+1) + \eta & P_i \in \mathcal{B} \\ \theta & P_i \in \mathcal{N} \\ \dfrac{\kappa - \theta}{1 - \epsilon - \alpha}(x - \alpha) + \theta & P_i \in \mathcal{G} \\ \dfrac{\kappa}{\epsilon}(1 - x - \epsilon) + \kappa & \mathcal{T}_i(p) > 1 - \epsilon \end{cases}$$

$$\ell_i = 0 \quad \Rightarrow \quad \mathcal{T}_i(p) = \mathcal{T}_i(p-1) - \mu'(x)$$

$$\mu'(x) = \begin{cases} \dfrac{\kappa}{\epsilon}(x+1) & \mathcal{T}_i(p) < \epsilon - 1 \\ \dfrac{\theta - \kappa}{\beta - \epsilon + 1}(x - \epsilon + 1) + \kappa & P_i \in \mathcal{B} \\ \theta & P_i \in \mathcal{N} \\ \dfrac{\eta - \theta}{1 - \alpha}(x - \alpha) + \theta & P_i \in \mathcal{G} \end{cases}$$

To ensure $\mathcal{T}_i(p-1) + \mu(x) \leq 1$ and $\mathcal{T}_i(p-1) - \mu'(x) \geq -1$ when $x = 1 - \epsilon$ and $x = \epsilon - 1$ respectively, $1 - \epsilon + \kappa \leq 1$ and $\epsilon - 1 - \kappa \geq -1$ must be satisfied, or equivalently $\kappa \leq \epsilon$. This is sufficient to ensure $\mathcal{T}_i(p)$ never exceeds $+1$ or -1.

C. Evaluation

The model is then evaluated from the following perspectives for further improvement: (A) *Behavioral*: how the model performs among a large enough number of players by running a number of standard tests, i.e., a sequence of "cooperation" and "defection" (or no-participation) for each player. Then, the result can be compared with the existing benchmarks or other models. (B) *Adversarial*: how vulnerable the model is to different attacks, e.g., the "Sybil attack" where the system is subverted by forging identities, or other kinds of corruption by a player or a coalition of malicious parties.

D. Modification

Adjustment might be required based on the evaluation results or system requirements. Let $\delta = \sum_{i=1}^{n} \ell_i$ denote the total number of cooperative players. Here we illustrate a sample modification of our specification: If $\delta = n$, i.e., all players have cooperated, it is not required to increase the trust value of anyone. If $\delta = 0$, i.e., all players have defected, it is not required to decrease the trust value of anyone. If $\delta > \frac{n}{2}$, i.e., majority of the players have cooperated, cooperation should be rewarded less and defection should be penalized more. If $\delta < \frac{n}{2}$, i.e., majority of the players have defected, defection should be penalized less and cooperation should be rewarded more. If $\delta = \frac{n}{2}$, i.e., the number of cooperative and non-cooperative players are equal, cooperation and defection should be readjusted with an equal ratio.

By using the same $\mu(x)$ for trust amplification and reduction, the modified function, termed *social trust function*, is as follows: $\mathcal{T}_i(p) = \mathcal{T}_i(p-1) + (\ell_i - \frac{\delta}{n})\mu(x)$. Note that this is just a sample revision to clarify our technical approach.

Conclusion

Our research attempts to bridge the gap between two literatures; proposes a novel and more interdisciplinary way of viewing the problem of trust and influence management; provides a valuable set of data; offers new hypotheses and computational models; and demonstrates how these new models can be used in technological systems. Furthermore, by mimicking human reasoning, the nature of trust and influence will become more apparent.

References

Castelfranchi, C., and Falcone, R. 1998. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *3rd International Conference on Multi Agent Systems*, 72–79. IEEE.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *Academy of Management Review* 20(3):709–734.

Mui, L.; Mohtashemi, M.; and Halberstadt, A. 2002. Notions of reputation in multi-agents systems: a review. In *1st ACM International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS'02*, 280–287.

Nojoumian, M., and Lethbridge, T. C. 2008. A new approach for the trust calculation in social networks. *E-business and Telecommunication Networks: 3rd International Conference on E-Business ICE-B* 9:64–77.

Nojoumian, M., and Stinson, D. R. 2012a. Social secret sharing in cloud computing using a new trust function. In *10th IEEE Annual International Conference on Privacy, Security and Trust, PST'12*, 161–167.

Nojoumian, M., and Stinson, D. R. 2012b. Socio-rational secret sharing as a new direction in rational cryptography. In *3rd Int. Conference on Decision and Game Theory for Security, GameSec'12*, volume 7638 of *LNCS*, 18–37. Springer.

Nojoumian, M.; Stinson, D. R.; and Grainger, M. 2010. Unconditionally secure social secret sharing scheme. *IET Information Security, Special Issue on Multi-Agent and Distributed Information Security* 4(4):202–211.

Nojoumian, M. 2012. *Novel Secret Sharing and Commitment Schemes for Cryptographic Applications*. Ph.D. Dissertation, Department of Computer Science, University of Waterloo, Canada.

Resnick, P.; Kuwabara, K.; Zeckhauser, R.; and Friedman, E. 2000. Reputation systems: facilitating trust in internet interactions. *Communications of the ACM* 43(12):45–48.