To be Others' Eyes and Ears?
- Trust and the Credibility of Reputation

Yen-Sheng Chiang

Department of Sociology,
University of Washington, Seattle, U.S.A

yen506@u.washington.edu

Introduction

The decision of trust is a frequently encountered issue for people of modern times. Buying used cars from strangers, donating money to charity organizations, and going to a new clinic are among the many examples where the placement of trust is involved. When making a decision, we rely on our past experiences and the information provided by others to evaluate the trustworthiness of the individuals we are dealing with. The latter mechanism includes formal channels, such as newspapers, and informal ones such as gossips spread in social networks.

For the mechanism of reputation to work effectively, a precondition needs to be satisfied that the spread of reputation be efficient and flawless. This assumption, unfortunately, does not always stand in reality. Not only does the spread of reputation have limited access, but it might become distorted and biased when being disseminated longer and further. A remedy to the problem is to introduce a new mechanism of reputation of a higher level—the reputation of the information of reputation. For instance, the website "Epinions" implements the idea where users not only can get access to the information of products or services, but they also know to what extent those raters' opinions reflect the truth.

The puzzle remains unsolved as some questions subsequently arise: who is going to evaluate the information of reputation? And on what basis, if any? The paper presents a simple model to address the issue. In particular, the whole system is treated as an interplay operating constantly between individual actions and collective information. Agents' trust decisions are guided by their memories and the information of reputation. The results of interactions are then fed back into the reputation system. In addition, agents re-evaluate the credibility of each information source in the reputation system by measuring the difference between the provided information and their own experiences. Agents learn to ignore dubious information sources and put more weight on the countable ones. When steady states are reached, both the module of trust actions and the credibility of information sources are in equilibrium—*individual actions in equilibrium will be guided (partially) by the in-equilibrium reputation system, which receives feedback from the former*.

Brief Literature Review

The issue of trust has been widely and intensively discussed.[1] One of the strengths that social simulation and agent-based models share uniquely is its capacity to model the dynamics of social process. There are many types of social simulations. The paper treats simulation as a modeling tool rather than a replication of real cases.

One type of social simulation on the dynamics of trust assumes that agents possess fixed strategies and undergo a selection process based on the accumulated payoffs derived from the

[1] For example, economists employed the "trust game" to articulate the trust dilemma (Kreps, 1990). Conducting laboratory experiments, social psychologists found that subjects behaved in a more trusting and trustworthy manner than rational choice theory would predict (Ostrom & Walker, 2003). For empirical studies, a great amount of work shows how interpersonal trust is embedded in social networks (Burt & Knez, 1995); how impersonal trust is related to trust in other social institutions such as the government (Rothstein, 2005), and how trust lubricates economic transactions when other formal mechanisms of enforcement are not in effect (Miller & Whitford, 2002).

repeated trust games. Macy and Sato (2002) and Bicchieri et al. (2004) are examples of this kind of work. The former treats learning as an adaptive mechanism while the latter employs the standard replicator equation in the evolutionary game theory to model the reproduction of agents. In their models, agents resort to their own memories of past interactions for guidance of actions. Agents do not "share" their memories, however. In the real world, the diffusion of reputation is commonly observed; for instance, many online auction websites, such as the "eBay", use the reputation system to sustain economic transactions (Kollock, 1999).

The reputation system suffers from numerous kinds of problems and biases, however. Paolucci (2000), and Conte and Paolucci (2002) discuss two kinds of errors that appear in the reputation system—either an honest agent is misjudged as corrupt one or a defector is mistreated as a countable one. Using simulations, Paolucci (2000) predicted how the different kinds of errors influence the prevalence of trust. Their prediction is pessimistic: when errors are introduced such that the reputation system is not effective, average welfare for honest agents is lower than cheaters. Whitby et al. (2004), on the other hand, propose a Bayesian reputation system to filter unfair evaluations. They show that when errors are not of great extent, the reputation system works well and gives accurate information of reputations.

The two studies mentioned above did not clearly address how agents would tell whether the information of reputation is true or false. Paolucci (2000) argued that agents would be more likely to disregard information released from cheaters. However, the problem is, trust decisions and spreading the information of reputations might be two independent matters: a cooperative (trustworthy) agent might be an uncountable messenger and vice versa. In the Bayesian reputation system proposed by Whitby et al. (2004), the so-called authentic opinion is actually a consensus of what most evaluators would agree on. As a consequence, if most opinions are false, then a single piece of true information will still be regarded as false simply because it deviates from the majority.

The model in this paper assumes that agents evaluate the credibility of reputations *by using their own experiences as reference points*. This is a simple and yet empirically supported assumption. After all, the most reliable way to assess the information of reputation is to experiment it by oneself. Admittedly, there are constraints to this approach. The following model parameterizes some key variables to show under what circumstances will trust and the information of reputation function well.

The Model

Consider a population of *N* agents. In each period they are randomly paired up. One of them is a trustor and the other trustee. The assignment of roles is randomly determined (probability= 1/2). The trustor makes the decision whether to trust the trustee or not. If s/he rejects to trust, the game is over. If trust is placed, the trustee decides whether to honor or abuse it. The setting is analogous to the trust game proposed by economists (Kreps, 1990). The logic of "backward-induction" in game theory argues that given the trustee's optimal choice is to abuse trust, the trustor will not trust in the beginning . Here, a stochastic dynamic model is proposed—agents' decisions are triggered probabilistically (Coleman, 1990; Gambetta, 1988), games are played repeatedly overtime, and the spread of reputation is considered. We argue that if trustors are exploited, they will be less likely to trust the same agent again. Trustees abstain from abusing trust when they realize that their bad reputation will rapidly be spread out to the public.

Let us suppose that initially every agent is fully trusting when being a trustor, and is tempted to abuse trust when being a trustee. It is our interest to observe how the tendencies of placing trust and honoring/abusing trust change overtime. There are two sources of information: individual memory of past interactions and the reputation system. *The weight on the two sources is a parameter to manipulate in the model*.

Each agent is assumed to remember interactions taking place in the past W_1 periods. In their memories, they remember whom they interacted with and what actions they took. Agents report the results to the reputation system which can be pictured as a central database that records agents' feedback.[2] The reputation system has a similar problem of memory decay, and it only records events in the past W_2 periods. Due to the stochastic nature of the model, reporting without errors does not mean the reputation system will provide accurate information all the time. A trustee might honor trust at one point in time, but abuse it at another even though the probability underlying the actions is the same. Conflicts between the reputation and self experience erode the credibility of the information source. On the contrary, consistency between them consolidates it.

Formally, let P_{ij} denote trustor *i*'s propensity to trust agent *j*:

$$p_{ij} = \begin{cases} 1 + \Delta p_{ij} & if \quad \Delta p_{ij} < 0 \\ 1 & else \end{cases} \tag{1}$$

where Δp_{ij} is the adjustment of trust propensity evoked by agent *i*'s memory of the history of interactions with *j* (m_{ij}), and *i*'s assessment of *j*'s reputation (r_{ij}):

$$\Delta p_{ij} = sm_{ij} \times (1 - s) r_{ij} \tag{2}$$

The parameter $s \in [0,1]$ controls the relative weight between the two inputs. The variables *m* and *r* are specified by the following two equations:

$$m_{ij} = \frac{1 \times M_{ij}^{+} + (-1) \times M_{ij}^{-}}{M_{ij}^{+} + M_{ij}^{-}} \tag{3}$$

where M_{ij}^{+} and M_{ij}^{-} represent, respectively, how many times in agent *i*'s memory *j* honors and abuses *i*'s trust. The records can be traced back to the past W_1 periods only.

$$r_{ij} = \sum_{k \neq i, j} E_{ik} R_{kj} \tag{4}$$

where E_{ik} refers to *i*'s assessment of the credibility of agent *k*'s evaluation of *j* (R_{kj}). Put in other words, r_{ij} is *i*'s perception of the reputation of *j* from different weighted sources of information. Records of R_{kj} in the reputation system are only saved for W_2 periods, and are calculated by the following equation:

$$R_{ij} = \frac{1 \times PM_{ij}^{+} + (-1) \times PM_{ij}^{-}}{PM_{ij}^{+} + PM_{ij}^{-}} \tag{5}$$

where PM_{ij}^{+} and PM_{ij}^{-} represent, respectively, the number of cases accumulated in the reputation system that agent *i* reports *j* honors and abuses his/her trust. To make a comparison, one can imagine equation (3) as agents' personal "bookkeeper" while (5) as a duplicate of the bookkeeper saved in the central database of the reputation system. The only difference is that, the lengths of time for the records to be maintained are different between the two.

For trustees, in the one-shot game their optimal choice is to abuse trust. When the game is repeated, they might worry that current abuse of trust will cause future losses. Agents assess the extent of this concern in two ways. First, they examine how many times they fail to earn the trust from others in their memory of interactions (d_j). Second, they check how effectively the reputation system operates by comparing her/his own experience *as a trustor* and what the reputation system informs

[2] For future studies, we can model that reputations are spread in fragmented social networks rather than being recorded in a central database.

her/him. By this definition, the effectiveness of the reputation system from agent *j*'s perspective (e_j) is thus $\sum_{k \neq j} E_{jk}$. Then *j*'s propensity to abuse trust (q_j) is:[3]

$$q_j = \begin{cases} 1 - \Delta q_j & if \ \Delta q_j > 0 \\ 1 & if \ \Delta q_j = 0 \end{cases} \tag{6}$$

where $\Delta q_j = s\,d_j + (1-s)\,e_j$ (7)

$$d_j = \frac{\sum_{i \neq j} M_{ij}^0}{\sum_{i \neq j} (M_{ij}^+ + M_{ij}^- + M_{ij}^0)} \tag{8}$$

$$e_j = \sum_{k \neq j} E_{jk} \tag{9}$$

M_{ij}^0 represents the number of times *i* fails to get the trust from *j*. Simply put, equation (7) to (9) argue that the more frequently trustees fail to gain the trust in the past, and the more effective the reputation system works, the less likely they will abuse trust. The same parameter, $s \in [0,1]$, controls the weight on self memory in contrast to the reputation system.

Finally, agents re-evaluate the credibility of information sources. They assign more weight on those who provide information similar to their own experiences. The following equation explicates the reevaluation process[4]:

$$E_{ij} = \begin{cases} \varepsilon + (1-\varepsilon)(1 - |m_{ik} - R_{jk}|) & if \ m_{ik} \times R_{jk} > 0 \\ \varepsilon \left(1 - \frac{|m_{ik} - R_{jk}|}{2}\right) & if \ m_{ik} \times R_{jk} < 0 \\ \varepsilon \in (0,1) \quad else & \end{cases} \tag{10}$$

where $\varepsilon = 0.1$ is the baseline credibility assigned to every agent. Equation (10) implements the idea that if an agent's experience is consistent with the information provided by another agent, i.e., they both agree the target agent is trustworthy or untrustworthy, the credibility of the information source will be confirmed, and the degree of consolidation is determined by the similarity between the two opinions. On the other hand, if an agent's experience is contradictory with the information provided by another, the former will put lighter weight on the latter's information provision next time.[5]

[3] Note trustee *j*'s decision is not agent-specific. Unlike the trustor who needs to decide whether to trust a particular agent in question, the trustee, whenever confronting a decision node, has already got the trust from the trustor. The only concern for her/him is whether the action she/he takes will be recorded in the reputation system or individual memory such that possible consequences would result to influence their future gains.

[4] In earlier periods of the simulation, there are not many records accumulated in individuals' histories. An extra constraint is imposed here such that agents will evaluate the credibility of the information only when they have their own experiences ready to be compared; that is, when both M_{ij}^+ and M_{ij}^- are non-empty sets.

[5] Note since $-1 \leq R_{jk}, m_{ik} \leq 1$, the maximum difference between the two variables will be 2 and this is why the denominator in the second condition of equation (10) is 2.

The pseudocode for the model is outlined as follows.

```
For each time period ~
        ■ Memory decays (for both individual memory and the reputation system)
        For each agent ~
                ■ Randomly paired up with another agent
                ■ Decides to be a trustor or trustee?
                ■ Trustor:
                     ● Given: $m_{ij}$ (individual memory of being exploited by j before)
                              $r_{ij}$ (reputation of j):
                     ● Trust or not?
                        If trusts ~
                        ■Trustee:
                                ● Given: $d_j$ (individual memory of not gaining trust before)
                                         $e_j$ (perception of the effectiveness of the reputation system)
                                ● Abuse or honor trust?
                ■ Record the outcome in individual's memory and the reputation system
                ■ Compare $m_{ij}$ and $R_{ij}$ to re-evaluate the credibility of each information source in the
                  reputation system
        end for agent
end for period
```

Results

A few representative graphs are selected for presentation. All simulations were run assuming a population size (*N*) of 100. In all graphs, the horizontal axis represents time periods and the vertical axis shows the proportion of each type of trust action, except for Figure 6 where the vertical axis records the effectiveness of the reputation system.

Notation Notes:
W1: the length of individual memory
***W2*:** the length of collective memory (reputation system)
s: the weight on self memory (**1-*s*** on the reputation system)

Legend:
Green color: the proportion of abusing trust
Blue color: the proportion of honoring trust
Red color: the proportion of trust not being given

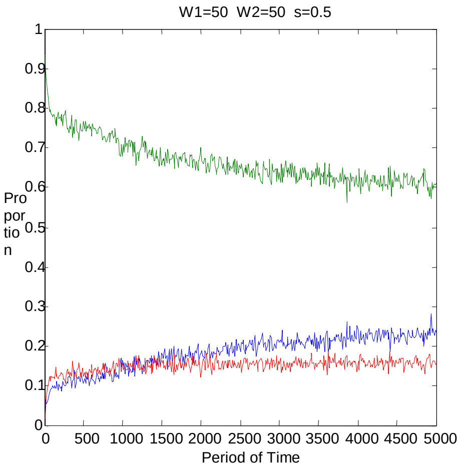

Figure 1

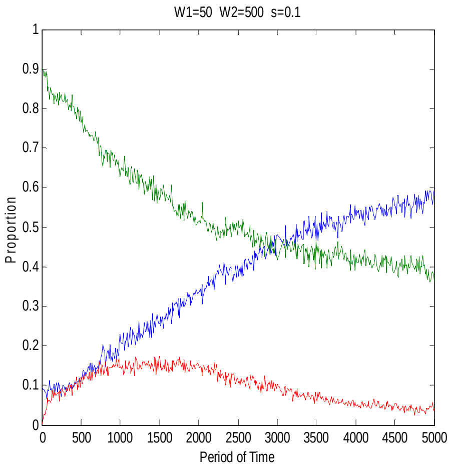

Figure 2

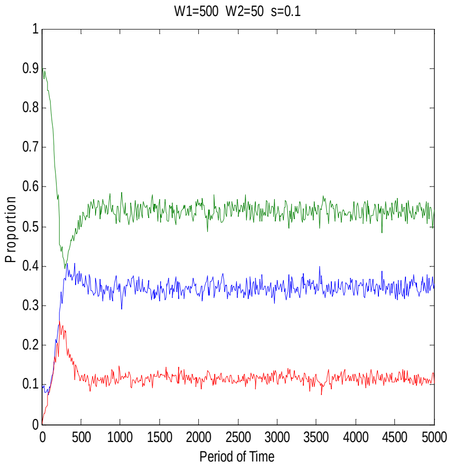

Figure 3

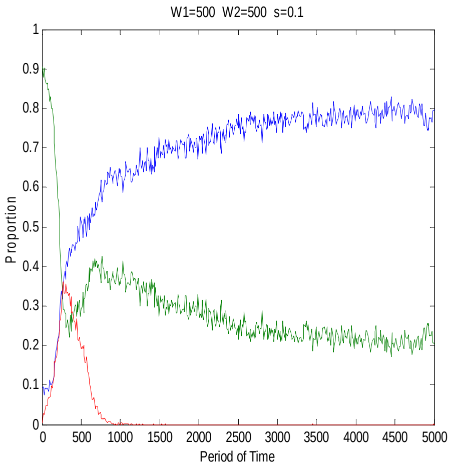

Figure 4

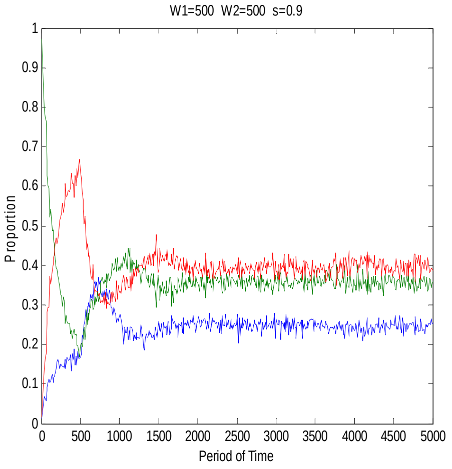

Figure 5

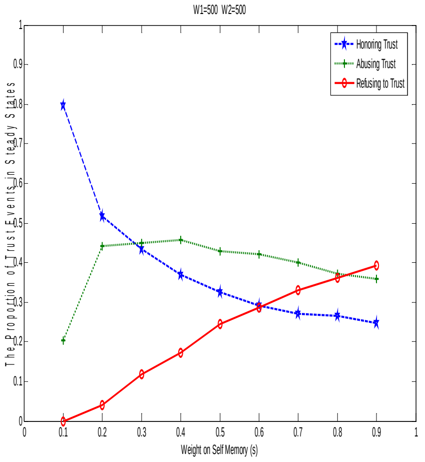

Figure 6

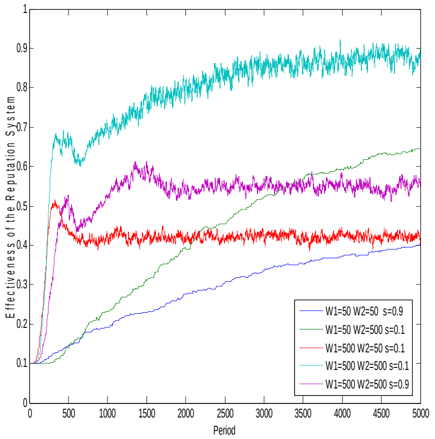

Figure 7

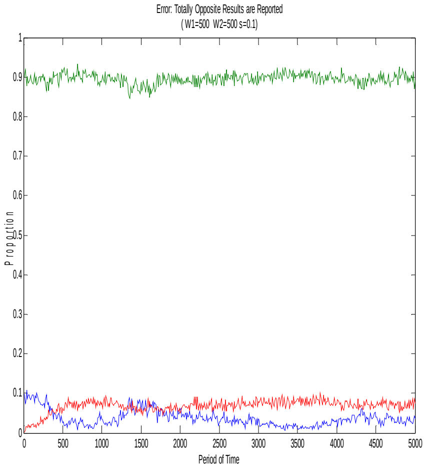

Figure 8

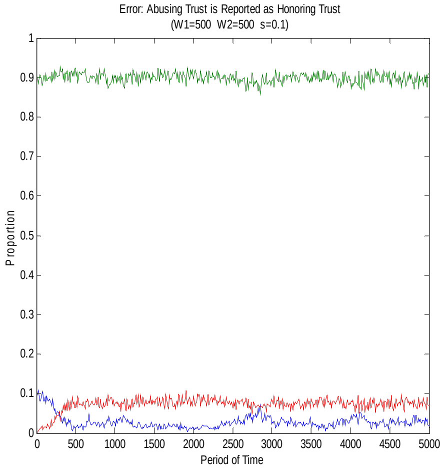

Figure 9

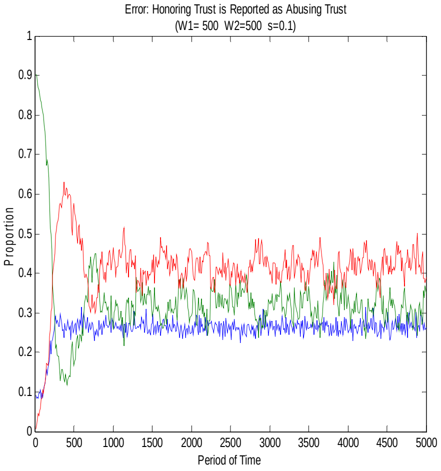

Figure 10

The figures exhibit the prevalence of the three types of trust events overtime. The frequency of abusing trust declines right after the dynamics embarks as individual memory starts to accumulate and the reputation system begins to work. It is also evidenced by the fact that more agents refuse to place trust. Honoring trust becomes dominant when individual and collective memories are kept fresh long enough, plus a heavier weight (1-*s*) on the reputation system (see Figure 4). If it is impossible to have well reserved records in both memory systems, having a longer memory in the reputation system seems to work better (compare Figure 2 and 3). The weight on the reputation system (1-*s*) is critical to the emergence of a high level of honoring trust (compare Figure 1 and 2, and Figure 4 and 5) as is shown explicitly in Figure 6 which records the proportion of trust events in steady states as *s* increases.[6]

We attempt to literally outline the dynamics of the simulation. As individual memory and the reputation system start to function, some trustors refuse to place trust. Witnessing this trend, some trustees start to withdraw from abusing trust. Moreover, as time passes individual memory and public records become more congruent between each other, and hence the effectiveness of the reputation system becomes stronger. Under the circumstance, trustees are further less inclined to abuse trust and the pervasiveness of honoring trust becomes self perpetuating. Whether honoring trust will be

[6] Due to stochastic factors, we average the proportion of each type of trust event in the last 100 periods before the simulation stops at time=5,000.

prevalent in the end is determined by the values of the parameters set in the model. Among them, the effectiveness of the reputation system is found to be the most important as can be observed in Figure 6. Sustaining an effective reputation system is a function of complex combinations of the lengths of individual and collective memories.

We can also examine situations where agents report their results to the reputation system with errors. Figure 7 presents the case where agents report totally the opposite results. Figure 8 shows the scenario where agents report an abusing-trust event as an honoring-trust one, but report honoring trust without errors. Figure 9 presents the opposite scenario to Figure 8. Since the introduction of errors weakens the effectiveness of the reputation system, trustees feel more secure to abuse trust. Figure 7 and 8 support the reasoning. However, when error is of the type that honoring trust is misjudged and spread as abusing trust, trustors will be discouraged to trust. We can see in Figure 9 that refusal of trust is more frequent than other types of events. More interestingly, abusing trust and refusing to trust form a series of "mirror images" along the horizontal axis. It implies whenever the degree of refusal to trust declines, trust abuse increases until the former is "awakened" to guard against the exploitation, but waning again when fewer events of abusing trust occur.

Future Direction

How exactly the reputation system works in reality in an empirical question. Before the advancement of digital technology, people in old times employ both centralized and decentralized means for the spread of reputation. Examples of the former include documents updated and saved in business organizations, such as the "guild" in the medieval ages, that record member's activities of economic transactions with other members. More often encountered is the latter mechanism such as gossips spread within a village. Even though databases in modern times can save a huge amount of information, the evaluations of the information of reputation could be fragmented; for example, the *networks* of trustworthy evaluators in the website "Epinions". An extension to the current model is expected to capture the situation where the reporting and drawing of the information of reputation is fragmented. That is, there would exist a network, either endogenously or exogenously determined, that designates who spreads the information to whom and who gets the information from whom. This extension definitely adds more complexity to the model, and yet is expected to yields richer results. The merit of the present model lies in the fact that, even built on the simple assumption of a centralized reputation system, intriguing results arise if we consider how agents verify the information by comparing their self experiences to the information.

References

Bicchieri C., Duffy, J. & Tolle, G. (2004). Trust among Strangers. *Philosophy of Science 71*: 286-319.

Burt, R., & Knez, M. (1995). Kinds of Third-Party Effects on Trust. *Rationality and Society* 7: 255-292

Coleman, J. (1990). *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Gambetta, D. (1988). *Trust: Making and Breaking Cooperative Relationships*. Cambridge: Blackwell.

Kollock, P. (1999). The Production of Trust in Online Markets. *Advances in Group Processes* 16.

Kreps, D. (1990). Corporate Culture and Economic Theory. In: *Perspectives on Positive Political Economy*, ed. Alt, K Shepsle. New York: Cambridge University Press.

Macy, M., & Sato, Y. (2002). Trust, Cooperation and Market Formation in the U.S. and Japan. *PNAS,* 99 (3): 7214-7220.

Miller, G., & Whitford, A. (2002). Trust and Incentives in Principal–Agent Negotiations. *Journal of Theoretical Politics* 14 (2): 231-267

Ostrom, E., & Walker, J. (2003). *Trust and Reciprocity.* New York: Russell Sage Foundation.

Conte, R., & Paolucci, M. (2002). *Reputation in Artificial Societies.* Kluwer Academic Publishers.

Paolucci, M. (2000). False Reputation in Social Control. In: *Applications of Simulation to Social Sciences* ed. Ballot, G. and Weisbuch, G. Stanmore: Hermes Science Pub.

Rothstein, B. (2005). *Social Traps and the Problem of Trust.* Cambridge: Cambridge University Press.

Whitby A., Josang, A., & Indulska, J. (2004). Filtering Out Unfair Ratings in Bayesian Reputation Systems. *Proceedings of the Workshop on Trust in Agent Societies* AAMAS