Trust Is in the Eye of the Beholder

Dimitri do B. DeFigueiredo, Earl T. Barr and S. Felix Wu*

Dept. of Computer Science – UC Davis

Revision November 2008

Abstract

We carefully investigate humanity's intuitive understanding of trust and extract from it fundamental properties that succinctly synthesize how trust works. From this detailed characterization we propose a formal, complete and intuitive definition of trust.

Using our new definition, we prove simple possibility and impossibility theorems that dispel common misconceptions, expose unexplored areas in the design of reputation systems and shed new light on the shortcomings of previous impossibility results.

1 Introduction

Many internet applications use trust to help users in their online interactions with others. The Feedback Forum, eBay's rating system, is perhaps the best known example. Trust, in different flavors, is particularly useful to distributed online applications such as peer-to-peer or social networks [10]. Both online and off-line, knowing how much to trust someone helps us know what to do in our interactions with them. To better understand the concept of trust online, we begin exploring mundane off-line real-world trust, the concept we seek to mimic.

Trust systems, such as the Feedback Forum, use reputation functions to quantify trust. Most of the trust systems deployed to date, use *consensus-based* reputation functions. Unfortunately, consensus-based reputation functions have intrinsic limitations, as we show in Theorem 4.1. In particular, they are exploitable — they can be manipulated by untrusted parties. We introduce a new class of reputation functions, *personalized* reputations functions, that overcome these limitations. Personalized reputation functions allow a person to control how much she trusts another party, independent of the opinion of others, thus allowing her to make her own, individual, assessment of each party's trustworthiness. Consensus-based reputation functions require both Alice and Bob to agree on how trustworthy Charlie is. Personalized reputation functions do not.

*{defigueiredo,etbarr,sfwu}@ucdavis.edu

To date, most trust systems use reputation functions that arbitrarily map trust to numbers. We show that trust is quantifiable in terms of utility. This insight links the trust transitivity problem, *i.e.*, how much should you trust the friend of a friend, to the problem of making interpersonal comparisons of utility.

People confuse trust and trustworthiness: they tend to think how much they trust a person is an objective measure of that person's trustworthiness, that everyone would agree with their trust assessment if only "they knew what I know." Trust is, in fact, independent of trustworthiness. Alice may trust Bob even though he is not trustworthy; Alice may distrust Charlie even though he *is* trustworthy. This confusion underlies the prevalence of trust systems built on consensus-based reputation functions. Below, we draw a bright line between trust and trustworthiness, setting the stage for trust systems built on personalized reputation functions.

We make the following contributions:

1. We introduce and formalize personalized reputation functions;
2. We provide a definition of Trust that exposes the link between trust transitivity and interpersonal comparisons of utility; and
3. We elucidate the difference between trust and trustworthiness.

This paper is structured as follows: In Section 2, we investigate what properties vying candidate definitions of trust should have in order to pick a definition that is as close as possible to our actual use of trust. We establish that trust is domain specific and explain the difference between trust and trustworthiness. Section 3 formalizes the intuitive understanding produced by Section 2. It also shows the link between the transitivity of trust and interpersonal comparisons of utility. We then use this formal framework in Section 4 to provide possibility and impossibility results. We show that all consensus-based reputation systems are exploitable by untrustworthy parties, but also that there are non-exploitable personalized reputation systems. The latter result dispels fallacies from previous impossibility theorems [3]. Section 5 discusses the implications of both results and shows how they relate to previous work. We present our conclusions in Section 6.

2 Characterizing Trust

What is trust? What is it used for? Different people give different answers to these questions. Like the word *love*, the word *trust* is overloaded with different meanings that convey different concepts depending on context. In addition to the simple semantic difference between a verb and a noun, each of the many different meanings of trust is compounded by nuance. To clarify the concept and think about it in a more disciplined manner, we need to agree upon a common definition of trust. Such a definition should capture as much of our intuitive understanding of the concept as possible.

We now examine our intuitive understanding of trust and try to extract from it fundamental properties that succinctly synthesize how trust works. First, we define our terms and note that trust is domain-specific, then establish the difference between a God's eye view of the trustworthiness of a person and human approximations of it. Both of these observations are contributions of this paper. We then turn to second-hand trust and trust transitivity, followed by the purpose and usefulness of trust. A person is necessarily vulnerable to the people she trusts, so we next discuss how a person decides how much to trust another. The discussion provides insights on how to monetize trust and why trust can grow with time. We close by arguing trust is not useful if a person does not control how much they trust another.

One important point we would like to make is that trust involves at least two roles. Creatively, we call this the *two-role rule*. These two roles are usually performed by two distinct parties. If I lend my car to my friend Bob, I **trust** that Bob will drive it safely. I am the **trusting** party. Bob is the **trusted** party, the person upon which I place my trust. There are cases where both roles, the **trusting** and **trusted** parties, are taken on by the same individual. For example, I know plenty of people who trust themselves to drive safely. But we argue that these degenerate cases do not present a violation of the *two-role rule*, only that the same party assumes both roles.

Another characteristic (or property) we generally associate with trust is that it is domain specific. In other words, we may trust a person in one domain, but not in another. For example, I trust my father to take care of my finances, but I do not trust him to be on time for dinner. We tend to aggregate disparate domains such as "being on time" or "handling money" into more general ones. In fact, I could say: "I trust my father", independent of domain. In these aggregate domains, we may not be able to decide whether we trust our friend more than our sister! These abstract domains may be partially ordered. However, if given a specific domain, we can usually provide a more specific assessment of trustworthiness. In what follows, we assume specific domains that are totally ordered.

There is an important distinction between trust and trustworthiness. It is possible that Alice trusts Bob even though Bob is not trustworthy. In other words, Bob is trusted but not trustworthy. It is also possible that Bob is trustworthy but not trusted. We see how much Alice trusts Bob as Alice's best estimate of how trustworthy Bob "truly is" and these two values do not necessarily match.

One particularly interesting aspect of trust is the way it propagates between different agents. If Alice trusts Bob and Bob trusts Charlie, how much should Alice trust Charlie? This is the *trust transitivity problem*. Clearly, trust is transitive to some degree. Many people use their friend's opinions about others to some extent. However, it is not clear that we all use the same rules to incorporate our friend's opinions into our own judgments. Whatever transitivity rules a person uses, the concept of trust people actually use allows others to use their friend's opinions. The definition of trust we adopt should allow an agent to employ whatever degree of transitivity they see fit, or at least allow each agent

to employ some transitivity.

2.1 Risk Exposure

What is trust used for? *We use trust to deal with risk — those risks that depend on the actions of others.* In a perfect world where we do not run any risks, we do not need to trust anybody: if there are no risks trust is not required. At another extreme, if everyone is completely trustworthy, there is no risk associated to the behavior of others.

For example, I completely trust my mother never to steal any money from me. If I am wrong and my mother does steal from me, my mother was not, in fact, completely trustworthy. Again we must conclude that the behavior of a completely trustworthy party does not present any risk. In some sense, this is everyone's personal definition of a *completely* trustworthy party: there is no risk associated to their behavior. Therefore, if everyone really is completely trustworthy there is no risk. The existence of trust, more specifically trustworthiness, eliminates or mitigates risk.

How much do we trust others? Trust reflects the risk of dealing with others. Generally, we avoid taking risks that depend on the behavior of people we do not trust. At the same time, we are willing to take risks that depend on the behavior of trusted parties. How much one trusts others limits how much risk one is willing to take. For example, you may be willing to lend a colleague money but not willing to give them your bank account's password. If Alice trusts Bob more than Charlie, then Alice is willing to risk more when dealing with Bob than with Charlie. Obviously, Alice may choose to behave however she pleases. But, by limiting her risk based on the extent of her trust, Alice limits her exposure. In other words, Alice tends to limit her exposure to Bob by how much she trusts him. Therefore, we observe that: *The amount of trust one places in another reflects the amount of risk one is willing to take in dealing with that person.*

There are many different sources of risk. We cannot eliminate all the risks we run in our lives. However, if we are willing to pay a price, we can eliminate or mitigate *specific* risks. For example, we can -*completely* eliminate the risk of falling to the ground in an airplane crash by never flying. Of course, this does not eliminate the different risk of having a jumbo jet crash in your backyard. We can also mitigate the risk of losing our home to a fire by having fire insurance. In this case, a fire may still occur, but the prize money will mitigate our financial losses. Another example would be to avoid trading in the stock market. We can *completely* avoid the risk of losing money on a company in the stock market if we do not buy its stock.

Similarly, we can completely eliminate the risk involved in dealing with most people simply by avoiding them. There are exceptions, because there are people whose actions unavoidably impact you. You are out of luck if you do not trust your President, as there is very little you can do to avoid dealing with the results of his actions. For most other individuals, however, we can avoid risk almost entirely by not dealing with them. In summary, we observe that: *Usually, how*

much risk I take when dealing with others is under my control, not theirs.

2.2 Non-exploitability

How do we control our risks? We control how much risk we take when dealing with others by limiting our exposure based on how much we trust them. *This is why trust is so useful.* For example, I am willing to give my car keys to a "semi-trusted" valet, but I will not leave a $100 bill in the glove compartment. In doing so, I limit how much risk I am willing to take. If I trusted the valet, I could also leave the $100 bill in the car. *How much trust I place on others acts as a constraint on my behavior* that prevents me from taking on too much risk.

Although trust is closely linked to the risk in dealing with others, there is an important difference between trust and risk. An individual may not have control over the risk he runs when dealing with others but he has *complete* control over "how much trust" he places on others, even though he may not be able to act upon it. The fact that you cannot escape from the consequences of your President's actions does not imply that you trust him. This is an important characteristic of trust that is usually implicit in our daily lives: *the trusting party controls "how much trust" is placed on the trusted party.* We call this property *non-exploitability.*

The trusting party's control over how much trust she places in the trusted party is independent of how many people ask to be trusted. Suppose I am buying a car. Ten car dealers may try to convince me that the gas guzzler is a good deal; however, if I do not trust them, I will choose not to run the risk and I will not be fooled. If one of the 10 car dealers is my brother, the situation changes because I trust my brother, independent of the will of the other 9 dealers. In other words, no matter how many car dealers try to convince me, if I do not trust them, I cannot be fooled. How much trust I place in others and, therefore, how much risk I take, is under my control. *No matter how many others collude to try to convince me otherwise.*

The trusting party does not change how much he trusts others because he cannot avoid the consequences of their actions. The President does not suddenly become more trustworthy only because we can no longer kick him out of office. Similarly, if a valet is parking my car, I do not suddenly trust him less just because I realize that I left a $100 dollar bill in my car. I will return to my car and retrieve the bill, but my course of action depends on how much I trust the valet and not the other way around. If, once I arrive at my car, the $100 dollar bill is still there, I may change how much I trust the valet. However, I will change my trust because I have more information about the valet, not because I could not have prevented the valet from taking the bill.

The previous example underscores the utility of trust. Trust *enables* the trusting party to control his risk exposure. The trust each individual, consciously or unconsciously, associates with others *enables* that individual (trusting party) to change his behavior to control how much risk he runs. Trust does not *require* individuals to change their behavior, but it rather *enables* them to change their behavior to control their risk exposure. In fact, we argue that *peo-*

ple use the concept of trust that enables them to avoid as much risk as possible. They do so because this is the most useful concept.

People can be wrong about how trustworthy someone is; people may be limited in what they can do, and therefore unable to avoid some risks; or they may not trust others and therefore be unwilling to take on risks, but everyone operates in terms of a concept of trust that maximizes their control over their risk exposure. Any definition of trust that cannot be used to minimize risk is incoherent, limiting and, in short, less useful.

Non-exploitability means that the trusting party can completely control how much trust she places on others. In the real world, the trusting party may not be able to completely control the risk, but she is always able to control the trust inside her head. Non-exploitable trust allows a trusting party to limit how much risk she takes on and, therefore, how much damage any collusion of malicious parties can do to her. Trust is not as useful if it is exploitable; because an exploitable definition of trust would itself limit how much the trusting party can control the risks she runs. Just as one would not learn to drive a car that you have to push around because it is not as useful as a conventional car, one would not use a concept of trust that can be exploited by others because it is not as useful as humanity's intuitive notion of trust. We argue that non-exploitability is fundamental trust.

Online, it is usually easy for a single party to issue themselves multiple identities (or pseudonyms). This "cheap pseudonym" [5] characteristic of most online applications enables a single malicious real individual to disguise himself as a group of distinct online identities. Thus, if the concept of trust people use is exploitable, when used online it will not only be exploitable by collusions of malicious individuals but also by a single individual who obtains multiple identities.

Is the concept of trust you use in your life exploitable? Do you feel more vulnerable purchasing goods online than at your local store? Does your use of trust *enable* you to change your online behavior to limit your online risk exposure however much you desire? If your answer to the last question is "yes", you probably instinctively use non-exploitable trust inside your head.

2.3 Building Trust

People build trust with time. The reason we do so is best explained with an example. Imagine that Bob wants to buy a pair of shoes. He has seen the shoes that he wants to buy at a local store for $100 and on eBay for $70. If Bob takes a chance and buys the shoes from the unknown seller at eBay and the transaction is successful, Bob will have saved $30 because he trusted the seller. Had the seller not been there, Bob could not have saved this money. If Bob repeats this transaction twice more he will save a total of $90. Thus, even if the seller at eBay does not deliver the shoes the fourth time Bob makes a purchase, and Bob loses $70, Bob is still $90-$70=$20 better off then he would have been if he had bought all the shoes from the trusted local store. In this sense, the fourth time Bob transacts with the eBay seller, he cannot lose even if the seller

is untrustworthy.

Every time the trusting party *chooses* to engage in a transaction where he has to trust someone, he does so because there is a reward. We assume that the trusting party will only engage in such transactions because the value of the reward more than offsets the risks taken. It makes no sense for the trusting party to choose to run those risks otherwise. The trusting party can then accumulate his gains over a sequence of successful transactions to insure future transactions with the same trusted party; this enables the trusting party to be more trusting with time [4].

2.4 Single Worldview Fallacy

Many people believe that how much they trust someone is an accurate objective measure of the trusted party. Therefore, the trusted party is assigned a label "trustworthy", or "not trustworthy". The interesting part is that this label becomes a characteristic of the trusted party only. In other words, if I believe my neighbor is not trustworthy, I believe that I do so because this is his nature. This lack of trustworthiness has nothing to do with me, I was only smart enough to observe it. Most people with some common sense would have been able to observe the same. Because of the prevalence of this belief, many assume that we all have a single true trust value and that differences in opinion are only due to lack of information about the trusted party.

For example, I trust my friend Bob and I believe that the only reason people do not trust Bob is because they do not know him well enough. If anyone knew Bob as well as I do, they would also trust him. To be more precise, many believe that each individual has a *single universal* trust value that everyone should use. Although appealing, we believe that there is no coherent way to assign a single universal number to each individual to serve as a reference for all trusting parties. We call this the *single worldview fallacy*.

For example, it does not matter how well Bob gets to know his mother-in-law; Bob will not trust her as much as his wife does. The difference in trust is not due to a lack of information in Bob's part, but because of the different relationship between the trusted and the trusting parties. Bob will never be his mother-in-law's daughter and, accordingly, his mother-in-law behaves differently toward him than she does toward his wife. How trustworthy Bob's mother-in-law is depends on who is asking the question. Therefore, we should not combine Bob's assessment of her trustworthiness with his wife's into a single global trust value. Two different values representing two different view points is better. In fact, we show in Section 4 that using two different values is not only consistent with the conventional non-exploitable concept of trust we humans have inside our heads, but essential to the real world application of trust.

3 Formalizing Trust

Let us now try to capture as much as possible of our intuitive understanding of trust in a formal framework. We start with a few definitions.

Definition 3.1 (trust values)**.** Trust (and trustworthiness) values are real numbers.

This definition implies that trust is quantifiable on a single scale and that, from a single agent's point of view, there is a total ordering when comparing how trustworthy different agents are. If Alice tells us that (for the same domain):

- Alice trusts Bob more than Charlie;
- Alice trusts Charlie more than Derek; and,
- Alice trusts Derek more than Bob.

We believe she would be willing to change her mind once we point out the inconsistency.

We do not make further restrictions such as assuming that trust is a binary variable with two possible values $(-1, 1)$ or that it is in the interval $[0, 1]$, except that we do require that all agents agree that higher values are better (see the definition of trust threshold below).

We would like to point out a useful interpretation for trust values. One can think of trust values as being specified in dollars. Positive values answer the question: How much money are you willing to risk on this person? The higher the value, the more trusted the person is. This leads to a similar interpretation for negative trusts values. If v_1 trusts v_2 a negative amount $-x$, then x specifies how much money v_1 would have to risk gaining (not losing) to be worth the risk of depending on the behavior of v_2. This interpretation motivates the definition of trust we provide later in this section.

Definition 3.2 (reputation graph)**.** A *reputation graph* is an annotated directed Graph $G = (V, E)$, where each vertex is an agent and each directed edge in E has an associated trust value. The reputation graph does not need to be complete.

We interpret this reputation graph G as follows. There is a directed edge labeled x from vertex v_1 to vertex v_2, if v_1 trusts v_2 the specified amount x. If there is no edge between two vertices, then the amount of trust between them is unspecified.

We could build such a graph by asking each agent about all others. For example, we could ask v_1: How much are you willing to bet that v_2 is a good cook? And then add an edge to the graph, starting at v_1 and ending at v_2, labeled by the corresponding dollar amount. If v_1 does not know how much he trusts v_2, we would not add an edge. In light of this interpretation, we consider all of a vertex' outgoing edges to be data local to and under the control of that vertex.

The next definition accomodates the transitivity of trust. Specifically, the degree to which people use second-hand opinions when considering a domain. For example, although I trust Bob to be honest, he is naive and therefore I do not trust his judgement about the honesty of others.

Definition 3.3 (world)**.** A *World* is a sequence of reputation graphs $W = \{G_k = (V, E_k),\ k \geq 0\}$, where all reputation graphs use the same set of vertices. We call the first reputation graph G_0 the *direct experience graph.* The following graphs are called the *1-indirect graph*, the *2-indirect graph*, and so on.

Let us use the domain of "being a good cook" as an example. When building the direct experience graph one should only use information obtained from actually eating the food cooked by another agent. No information received indirectly from other parties should be used. There should only be an edge going from v_1 to v_2 if v_1 actually tasted food prepared by v_2 and has an opinion on it. Similarly, the 1-indirect graph should answer the question: Is this person a good food critic? In other words, Does v_1 trust v_2's palate to evaluate someone else's cooking? The sequence of graphs represents increasing levels of indirection for the same initial domain.

A world has a single set of vertices, but otherwise the graphs G_k can change arbitrarily for different k. Each graph has a different set of edges E_k, each labeled with different trust values. The definition of a world mirrors the multiple domains of trust and that there may not be correlations between these domains. Our results hold regardless of whether or not there exist correlations between edges in different graphs: the definition takes into account settings where some agents trust their cooks as good food critics and other agents do not.

Definition 3.4 (reputation function)**.** Given a world, a *reputation function* f assigns a trust value to each ordered pair of vertices.

Definition 3.5 (trust graph)**.** A reputation function outputs a *complete,* directed *trust graph.*

The reputation function tells us how trustworthy v_1 thinks v_2 is, for every pair (v_1, v_2) in the world. Because two parties may not have interacted in the past, we let a reputation graph be incomplete. However, because an agent does not know whom it will meet next and may need to suddenly form an opinion, we require the trust graph to be complete. Trust graphs and reputation graphs represent different concepts. The reputation graphs represent the reputation information available. Given that information, the trust graph expresses how trustworthy each party is, from different points-of-view. We will consider the trust graph in detail later. Given the world W, we use $f[W](v_i, v_j)$ to denote the trust value f assigns to the edge (v_i, v_j) in the trust graph. When W is clear from context, we simply use $f(v_i, v_j)$.

We require that all the reputation information available to a reputation function be expressed in the edges and edge labels of the different graphs that constitute a world. No other information should be used. This can be made precise by the following isomorphism requirement:

Definition 3.6 (normalization)**.** Let us denote by πW the world obtained by permuting the vertices of W with permutation $\pi : V \rightarrow V$. A *normalized reputation function* f is one where $f[W](v_i, v_j) = f[\pi W](\pi(v_i), \pi(v_j))$ for any permutation of the vertices π and for any world W.

The above definition simply states that if we permute the input to the reputation function we should obtain the same permutation of the output. This requirement forces reputation functions not to obtain extra information from their input in the form of vertex labels. For example, assume that v_2 trusts v_1 because v_1 represents the *New York Times.* In other words, v_1 has the vertex label *"New York Times"*. Why does v_2 trust the *New York Times*? Is it not because of the New York Time's *reputation*? If so, this reputation information should be added to the world as edges (v_2, v_1) with the corresponding edge labels. Note that it is easy to transform vertex labels into edges that convey the same information.

Normalization forces any knowledge about the world to be made explicit in the edges and edge labels of the world given as input to the reputation function. This requirement also prevents reputation functions that perform better only when there are distinguished nodes from being mistaken for functions that work well in more general settings. It does not prevent such functions from obtaining information, it just makes the information used explicit. The normalization helps prevent the designer from comparing apples to oranges when comparing two reputation functions. For the remainder of this paper, we put all reputation information on edge labels and will restrict our attention to normalized reputation functions.

Definition 3.7 (consensus-based vs. personalized)**.** A *consensus-based reputation function* is one where, for all vertices v_j in the world graph, the trust values x_{ij} assigned to ordered pairs of vertices (v_i, v_j) are the same for all pairs ending in the same vertex v_j. Therefore, the trust value $x_{1j} = x_{2j} = ... = x_{nj} = y_j$ assigned by the reputation function is completely determined by v_j and denotes v_j's trustworthiness to all other agents. A reputation function that is not consensus-based is *personalized*.

In effect, a consensus-based reputation function assigns a trust value to each vertex in the trust graph it outputs, instead of assigning a trust value to each edge in that same graph.

Definition 3.8 (trust threshold)**.** The *trust threshold* for agent v_i is a trust value h_i established by that agent. All agents v_j whose trust values for (v_i, v_j) assigned by the reputation function are above the threshold h_i are **trusted** by v_i; otherwise they are **untrusted**. In other words, if $f(v_i, v_j) \geq h_i$ then v_i trusts v_j; otherwise v_i does *not* trust v_j.

Note that this classification of agents as trusted and untrusted is very flexible. It allows agents with negative trust values to be trusted or agents with positive trust values to be untrusted depending on the trust threshold each trusting party chooses.

We can now also define what we mean by distrust. In everyday life, we can usually classify others as trusted, neutral or distrusted. This leads to a simple definition of distrust: all agents strictly below the trust threshold are distrusted; *i.e.*, if $f(v_i, v_j) < h_i$ then v_i distrusts v_j. This is a very simple notion, but we have already captured it with our definition of an untrusted agent above.

We believe there are two alternative, equally intuitive, possible definitions of this concept. One is that an agent v_j is distrusted by the trusting party v_i if $f(v_i, v_j) < 0$. According to our previously suggested interpretation of trust values, this means that a distrusted party is one for which the trusting party needs to be tempted by the possibility of gain in order for him to interact with the trusted party. Another possible definition is that *a node is distrusted whenever it is trusted less than a complete stranger* [1]. This definition implicitly assumes that all complete strangers are trusted to the same extent by the trusting party v_i. We call this the *stranger threshold* s_i.

Most of time the stranger threshold is zero and the definitions are equivalent and equally appealing. However, it is possible for one to be trusting of strangers and thus willing to take risks that depend on their behavior. This would be analogous to setting a high stranger threshold, *i.e.* $s_i > 0$. This does not necessarily mean that strangers are trusted (as that requires that $s_i \geq h_i$ and it may be that $0 < s_i < h_i$), but only that the trusting party is willing to take some risks based on their behavior.

We adopt the latter definition for distrust. We do so because it brings out the observation that distrust is useful, mostly when parties cannot easily get new identities. Otherwise, it is easy for malicious players to issue themselves a new identity everytime they want to fool somebody. The "cheap pseudonym problem" [5] rears its head again. We are now ready to provide a definition of trust.

Definition 3.9 (Trust)**.** *Trust* is the personal threshold determined by the trusting party that describes the maximum utility the trusting party is willing to risk when dealing with the trusted party.

The definition quantifies trust in terms of utility and is a significant contribution of this paper. Expressing trust in terms of utility makes the link between the trust transitivity problem and interpersonal comparisons of utility very clear. The trust transitivity problem is a consequence of and reduces to the problem of making interpersonal comparisons of utility. The last two definitions also make precise colloquial questions like: "How much do you trust him?" and "Is he trustworthy?"

Definition 3.10 (collusion)**.** A *collusion* is a subset of the vertices of a World.

Definition 3.11 (untrusted collusion)**.** An *untrusted collusion*, from agent v's perspective, is a collusion whose members are *all* untrusted agents.

Definition 3.12 (manipulated world)**.** Given a world $W = \{(V, E_k)\}$ and a collusion $C \subseteq V$, a *world manipulated by that collusion* $W'_C = \{(V', E'_k)\}$ is a

[1] For brevity we omit the formal definition of complete strangers.

modified world in which the collusion can change the reputation graphs in two ways:

- The collusion can arbitrarily add or remove edges starting from any member of the collusion; *i.e.,* for any $v_i \in C$ and $v_j \in V$, we can add or remove any edge of the form (v_i, v_j) in any reputation graph.
- The collusion can arbitrarily change the trust values on any edge starting from any member of the collusion (including any edges they have added).

In all other respects W'_C is identical to W. In particular, $V' = V$.

A non-exploitable reputation function is one for which each trusting party can determine autonomously and arbitrarily how much she trusts others. This implies that no untrusted collusion can increase the trust value of any agent. This enables the trusting party to control how much trust she places on others and to set an absolute bound on how much damage any collusion of malicious agents can do. The trusting party can still take risks if she so desires.

Definition 3.13 (non-exploitability)**.** A *non-exploitable reputation function* is a reputation function where: For any given world W, *for any vertex* v_i and any trust threshold h chosen by that vertex, no **untrusted** collusion in W can change the trust value of any agent v_j, *i.e.*$f(v_i, v_j)$. This is done by comparing trust values in two worlds: the standard honest world W and *any* world manipulated by an untrusted collusion.

This formalizes the concept developed in section 2.2. Notice that this definition has six quantifications. We quantify over all worlds, for each world over all agents (trusting party role) and any trust threshold that agent may choose. Then, for each agent, for all possible collusions. For each such collusion, we consider all possible manipulations due to that collusion. Finally, for each possible manipulation we look at all agents (trusted party role) and ask whether that manipulation was able to change that agent's trust value. The quantifications are:

- over all worlds
- over all trusting parties and their thresholds
- over all collusions
- over all manipulations that collusion can perform
- over all agents whose values may be affected by these manipulations.

This definition encompasses *whitewashing* attacks [6], where an agent sheds a bad reputation and reappears as a newcomer, by its first quantification over all worlds. This quantification includes a world where both the new and old identities exist and collude with each other. Similarly, *sybil* attacks [3], where a malicious agent obtains multiple identities, are modelled as collusions.

A trivial reputation function is one that completely ignores the transitivity in trust and only takes into account local data.

Definition 3.14 (triviality)**.** A *trivial reputation function* f is a reputation function such that for any given world and for any pair (v_i, v_j) the value of $f(v_i, v_j)$ depends only on edges starting at v_i and is independent of all other edges.

This definition is actually slightly more general than required by our proofs. We are only concerned about ruling out a reputation function that is fixed. We could have opted for a more restrictive definition as in [3] or [2]. However, we believe that triviality also applies to reputation functions that use only local, but not necessarily fixed, information. That being so, triviality is better defined as above.

4 The Limitations of Trust

Our impossibility result follows clearly from the definitions above.

Theorem 4.1. *All non-exploitable consensus-based reputation functions are trivial.*

Proof. We assume that a consensus-based reputation function f is non-trivial and show that it is exploitable.

If there is only one node, all information is local and all reputation functions satisfy the condition for triviality, therefore we will assume that there exist at least two nodes. Assuming we have a reputation function that is consensus-based, any trusted party has a single universal trust value. Consider any two distinct nodes v_1 and v_2. We define two untrusted collusions — U_1 (untrusted by v_1), and U_2 (untrusted by v_2). The set of all nodes is V and $\overline{U}_1 = V - U_1$ is the set of all nodes not in U_1. In the trust graph output by the consensus-based reputation function f using the honest world W as input, v_1 has a trust value of t_1 and v_2 has a trust value of t_2.

Assume that U_1 now wants to trick the trusting party v_1 and change t_2. There are only two possibilities: either U_1 is able to change t_2 or not. If U_1 can change t_2, then f is exploitable, by definition. We therefore continue with the assumption that U_1 cannot change t_2, and consider the situation from v_2's perspective and ask "Can $\overline{U}_1$ change t_2?"

As with U_1, $\overline{U}_1$ can either change t_2 or not. If $\overline{U}_1$ cannot, f is trivial, since, if neither U_1 nor $\overline{U}_1$ can change t_2, t_2 is fixed (whomever v_2 might be) and we have contradicted our assumption that f is nontrivial. We therefore continue with the assumption that $\overline{U}_1$ can change t_2, and show that there exists a world in which $\overline{U}_1 \subseteq U_2$.

If $v_2 \in U_1$, it cannot change t_2, since we have assumed U_1 cannot change t_2. If $v_2 \notin U_1$ and it can change t_2, this fact does not make v_1 exploitable since v_1 trusts v_2: the question is whether v_2 can be exploited. Regardless of whether $v_2 \in U_1$ or not, v_2 can change h_2, its trust threshold. In particular, v_2 can set h_2 to be greater than the greatest trust value of any node in V. In so doing, v_2 sets $U_2 = V$ and no node is trusted. Thus, $\overline{U}_1$ is an untrusted collusion in v_2's eyes and f is exploitable against v_2.

So we have shown f is exploitable against v_1, or, if it is not, f either is trivial, in contradiction of our assumption that f is nontrivial, or f is exploitable against v_2. Therefore, f is exploitable. □

This theorem does not limit all reputation functions, only consensus-based ones. However, it does show that all non-trivial consensus-based reputation functions are exploitable. In other words, consensus-based reputation systems are intrinsically insecure. There is always a way to manipulate such systems.

Theorem 4.2. *There are non-trivial non-exploitable personalized reputation functions.*

The proof that follows is constructive. We build a binary (trusted/untrusted) reputation function that expresses a simple transitivity notion: "I trust you, if I trust somebody who trusts you". Without loss of generality, we set a global "reputation threshold" λ, but this is distinct from and does not preclude agents from picking their own trust thresholds. However, the binary nature of this toy function f means that trust thresholds outside the interval $(0, 1)$ lead to either everyone being trusted or untrusted, whereas all threshold settings within the interval are mutually indistinguishable. The function f also assumes that all reputation graphs are identical and disregards all but the direct experience graph. This is not as unrealistic an assumption as it may seem at first: if you are a good cook you can probably identify other good cooks. We proposed a reputation system based on a non-trivial, non-exploitable reputation function in previous work [4]. We refer the reader to that text for a complete description of a non-binary reputation function that monetizes trust values.

Proof. From lemmas 1 and 2 below we obtain that the function f_λ is non-trivial and non-exploitable. □

Definition 4.1. Given a world W, there is a *trusted path* from vertex v_i to vertex v_j, if there is a directed path in the direct experience graph from v_i to v_j along which all edges are labeled by at least λ.

Definition 4.2. Given a world W, we say that vertex v_i can reach vertex v_j if there is a trusted path from v_i to v_j.

Define f_λ as follows:

$$f_\lambda(v_i, v_j) = \begin{cases} 1, \text{ if } v_j \text{ can be reached from } v_i \\ 0, \text{ otherwise} \end{cases}$$

Lemma 4.3. *The function f_λ is non-trivial.*

Proof. Let us set *w.l.o.g.* the reputation threshold $\lambda = 0.5$. Consider the direct experience graph with four nodes v_1,v_2, v_3 and v_4 depicted in Figure 1. The node v_4 has no incoming or outgoing edges; whereas v_3 can be reached from v_1 through the trusted path $(v_1, v_2), (v_2, v_3)$. The only asymmetry between v_3 and v_4 in the graph is that there is no edge (v_2, v_4) above the threshold λ while the

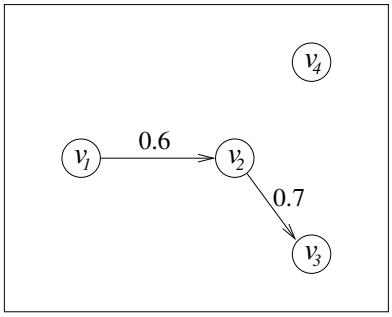

Figure 1: An example direct experience graph.

edge (v_2, v_3) exists and is above the threshold. This data is not local to v_1 (it is local to v_2), but, because of the data at v_2, f_λ outputs different trust values making $f(v_1, v_3) = 1$ whereas $f(v_1, v_4) = 0$. Therefore f_λ depends on non-local data and is non-trivial. □

Lemma 4.4. *The function f_λ is non-exploitable.*

Proof. Consider a node v_i. If v_i's trust threshold is outside the interval $(0, 1)$ then either all nodes are trusted or untrusted. If all nodes are trusted, there can be no untrusted collusion and f_λ is (vacuously) non-exploitable. Similarly, if v_i's trust threshold is larger than one, then all nodes are untrusted and no collusion can change that outcome. If either holds for any node v_i, f_λ is non-exploitable.

On the other hand, if v_i's trust threshold is within $(0, 1)$ then v_i trusts a node v_j if there is a trusted path from v_i to v_j. Because untrusted nodes cannot be reached, they cannot make other nodes reachable. Similarly, because untrusted nodes are never in a trusted path to any trusted node (otherwise they would themselves be trusted), they cannot make a node unreachable either. From v_i's perspective, no untrusted collusion can change trust values and this is true for any node v_i. Therefore, f_λ is non-exploitable. □

Take eBay's reputation system as an example. It uses a single global trust value for each party and is therefore consensus-based. Our impossibility theorem shows that, because of this limitation, the system can be manipulated by untrustworthy parties. At the same time our possibility theorem shows that, if the ratings were displayed only after a user logged in, the system could be personalized for that user and overcome this limitation. There is no way to completely secure a consensus-based reputation system and eBay's system, as it stands, will always be vulnerable to malicious manipulations. For a non-trivial system, such as eBay's, the only way to prevent manipulation is to personalize the system to each user.

5 Discussion and Related Work

Reputation systems [13, 12] and the related topic of "trust management" have received a lot of attention from the computer science community. A good survey on trust from this point of view can be found at [10].

There is a close relationship between reputation functions and reputation systems [13]. To be of practical use, reputation systems must aggregate information about each agent's past history into an easy-to-use format. It is usually easier to perform a *single* aggregation for all users. However, this is not a requirement. The view that producing a single set of aggregation results for all users is the only way to aggregate the past history is precisely the *single worldview fallacy*. We do not imply that a single global worldview is not useful; to the contrary, it is very useful. However, it has limitations that can be overcome if one personalizes the results obtained (See Definition 3.7). Furthermore, personalizing the results obtained does not mean discarding information. Ranking systems such as the ones considered in [2] can be changed to provide personalized results for many e-commerce applications and multi-agent systems. There need not be a single ranking of alternatives, each party should be allowed to have their own preferences.

Non-exploitability just means that the trusting party is able to arbitrarily control how much trust she places on others, no matter what world setting she faces. As seen from section 2.2, this is an important characteristic of real-world trust that is captured by non-exploitable reputation functions. Non-exploitable reputation functions and systems output personalized worldviews, one view for each agent. Non-exploitability is what makes them more useful and easier to understand. How would you use trust if you could not control how much trust you could place in others? This important property has been overlooked by the literature.

Our definition of normalization (Definition 3.6) does not limit our characterization of trust. Cheng and Friedman partition reputation functions into symmetric, those whose output is solely dependent on edges, or actual interactions, and asymmetric, those whose output is computed with respect to a distinguished node [3]. The isomorphism property we use to define normalization is analogous to the isomorphism property presented in [2] and the definition of a symmetric reputation function. Symmetric functions are subject to Cheng and Friedman's impossibility result. Asymmetric functions rely on nodes labels that encode extrinsic information. In other words, asymmetric functions correctly encode trust only if they were correctly constructed from previously known trust information. This is a form of the chicken and egg problem. Normalization avoids precisely this problem. The importance of normalization should not be played down because it seems too restrictive [3]: it is an essential requirement.

Reputation functions that are not normalized can be easily mapped to normalized functions. Our results hold given the appropriate mapping of the different domains. Furthermore, our possibility theorem contradicts previous results. Specifically, Cheng and Friedman show that "There is no symmetric sybilproof nontrivial reputation function" because they did not consider per-

sonalized reputation functions [3]. Here, we show that there do exist normalized non-exploitable nontrivial *personalized* reputation functions. Elsewhere, we have shown that it is possible to build a practical trust system from such reputation functions [4]. Cheng and Friedman's misconception would be hard to dispel without the intuitive understanding of trust developed in this paper.

There is a large body of literature on trust from a variety of disciplines. Alternative definitions in a computational setting have been proposed for both trust [11] and distrust [8], and an interesting attempt to classify all the different points of view can be found in [1]. Our definition is closely related to the one found in the encompassing work of Gambetta [7]. However, our definition not only observes the link between trust and risks that depends on the behavior of others, but also makes clear the fundamental connection between trust and utility. This link casts light in the trust transitivity problem and shows how trust can be monetized.

Further testament to the usefulness of the formal definitions proposed comes from their breadth. Our definition of a world provides a very flexible trust transitivity framework and encompasses many of the settings found in the trust propagation (or transitivity) literature [4, 9, 8, 14]. Despite the opposing results, we feel that the aims of our work most closely resemble those of [2] and [3].

6 Conclusion

Humans have an intuitive understanding of trust and are very proficient at using it as a tool to help them in their daily interactions with others. To make online interactions easier for users, a number of internet websites and peer-to-peer networks provide systems that explicitly attempt to capture the concept of trust. However, many of the systems in use today assign a single universal trust rating to each participating party. This implies that these systems are inherently vulnerable to manipulation by malicious users and are, therefore, not as useful to the end users as they could be.

Non-exploitable systems that do not provide a single universal trust rating, but that can change the trust ratings assigned to an individual, depending on domain and on who is asking the question, resist malicious manipulation. These systems are inherently more intuitive than the systems currently in use, and will become more useful tools to end users who seek help in dealing with the online world.

References

[1] http://www.istc.cnr.it/T3/map/index.html

[2] A. Altman and M. Tennenholtz, "Incentive Compatible Ranking Systems", *Proc. of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems.* (AAMAS'07), 2007.

[3] A. Cheng and E. Friedman, "Sybilproof Reputation Mechanisms", *SIGCOMM'05 Workshops*, Aug 2005, ACM.

[4] D. do B. DeFigueiredo and E. T. Barr, "TrustDavis: A Non-Exploitable Online Reputation System", *In Proceedings of Seventh IEEE International Conference on E-Commerce Technology* (CEC'05), 2005.

[5] E. Friedman and P. Resnick, "The Social Cost of Cheap Pseudonyms", *Journal of Economics & Management Strategy*, vol. 10, n 2, 2001, pp 173 - 199.

[6] E. Friedman, P. Resnick and R. Sami, "Manipulation-Resistant Reputation Systems", Chap. 27 of Nisan *et. al.* (ed), *Algorithmic Game Theory*, 2007, Cambridge Press, pp 677 - 697.

[7] D. Gambetta, "Can we Trust Trust?", Chap. 12 of Gambetta (ed) *Trust*, 1990, pp 213-237, Blackwell.

[8] R. Guha, R. Kumar, P. Raghavan and A. Tomkins, "Propagation of Trust and Distrust", in *Proceedings of the Thirteenth International World Wide Web Conference* (WWW2004), 2004, pp 17 - 22.

[9] S. D. Kamvar, M. T. Schlosser and H. Garcia-Molina, "The EigenTrust Algorithm for Reputation Management in P2P Networks", in *Proceedings of the Twelfth International World Wide Web Conference* (WWW2003), *ACM, 2003,* pp 20 - 24 .

[10] H. Li and M. Singhal, "Trust Management in Distributed Systems", *IEEE Computer Magazine*, vol 40, issue 2, Feb 2007, pp 45 - 53.

[11] S. P. Marsh, "Formalising Trust as a Computational Concept", *Ph. D. Thesis*, University of Stirling, April 1994.

[12] S. Marti, "Trust and Reputation in Peer-to-Peer Networks", *Ph.D. Thesis*, Stanford University, May 2005.

[13] P. Resnick, R. Zeckhauser, E. Friedman and K. Kuwabara, "Reputation Systems", *Communications of the ACM*, vol. 43, issue 12, Dec 2000, pp 45 - 48.

[14] T. Riggs and R. Wilensky, "An Algorithm for Automated Rating of Reviewers", in *Proceedings of the First ACM/IEEE-CS joint conference on Digital libraries*, 2001.