



# Trust Dynamics

Norman Foo

Computer Science and Engineering

University of New South Wales, Australia

*and* National ICT Australia

July 1, 2009

# Outline of Talk

1. Trust – we all understand it intuitively
2. A Crisis of Trust Breakdown
3. Interactions, Games, Norms, etc.
4. Cooperation
5. Experience, Trust and Reputation

## Trust – Our Intuitions

“I don’t trust him (her)!” How often have you heard that, or even thought that to yourself? About a relative, a friend, a politician, a banker, etc.

“People do not trust that company (party, radio station, country, etc.).” We hear a lot of that these days.

When you buy or sell something on eBay, you may rate the other party. What do the aggregated scores mean? Trustworthiness? Reputation? Are these the same?

You lend money (car, apartment, etc.) to someone. Trust is obviously involved.

“The Emperor has lost the Mandate of Heaven.”

## A Crisis of Trust Breakdown

The current international finance crisis has affected everyone.

Ralph Norris, CEO of the Commonwealth Bank (Australia), believes the cost of credit has risen because the market now has a whole new *perception of risk*.

“In fact, it’s basically been a *breakdown in trust*,” says Norris, whose bank is one of the most exposed to the highly publicised troubles of corporate Australia.

Banks refuse to lend to other banks because they *not trust* the latter’s published balance sheets. Banks refuse to lend to businesses for the same reason. Finance is paralysed. Businesses go

bankrupt. People lose jobs and have no money to spend. The cycle spirals downwards.

All because the trust that was once there is gone. Why? How?

Can trust be modeled? What affects trust? let us now look at a very simple scenario.

## The Classic Prisoners' Dilemma (PD)

Player 2

		C2	F2
Player 1	C1	(2,2)	(0,3)
	F1	(3,0)	(1,1)

Two partners in crime are interrogated separately. “C” means they cooperate with each other by refusing to admit; “F” means they fink by betraying the other party. If they cannot communicate or trust each other, they end up in the right bottom of the normal form matrix. *Lack of trust is the problem for this bad Nash equilibrium.*

## Nash Equilibrium

A central concept in game theory and multiagent interaction. Consider two players (agents) in a perfect information game defined by a payoff matrix. Assume both players know this matrix — that implies rational, mutual reasoning.

A pair  $(m_1, m_2)$  of “moves” is a Nash equilibrium if whenever player 1 chooses  $m_1$ , player 2’s best response (maximizes his payoff) is by choosing  $m_2$ , and vice-versa. It is a fixed point. There may be more than one such equilibrium in a finite game. Nash’s basic result is that there is always an equilibrium, which may be probabilistic.

It is not in the rational interest of either party to deviate from a Nash equilibrium (if there is no enforceable contract).

## Iteration

In the early '80s a University of Michigan professor Robert Axelrod configured a tournament in which people were invited to submit programs to play repeated versions of the PD.

Programs were played pairwise repeatedly and payoffs summed. This is the *Iterated PD*.

The winner was a strategy submitted by Anatol Rapoport called Tit-for-Tat. It starts by cooperating on its first encounter with another program, then copies that program's previous move thereafter.

Tit-for-Tat thus assumes a nice opponent on introduction, and thereafter it punishes at an encounter if and only if the opponent was bad the previous encounter. If the opponent becomes nice, Tit-for-Tat reciprocates.

If two players use Tit-for-Tat, they end up cooperating and in the top left of the matrix repeatedly and do very well.

Unfortunately, it can be shown that repeated betrayal by both players is also a “stable” state, an equilibrium – thus the cycle of revenge one sees in Iraq and Pakistan.

Can we find conditions to avoid bad equilibria?

# The Emergence of Cooperation

Since that time many researchers have investigated this emergence of cooperation in settings where individual encounters with no prospect of repeat encounters seem to doom the parties to bad behavior.

The key ingredient appears to be *unpredictable iteration*. This is intuitive: If you do not know when you have to deal with someone again and again, it is wise to be nice this time — that is the bottom line. At a higher level if you have reason to trust someone, being nice is prudent so that he/she will not lose trust in you in future transactions.

## A Closer Look at the Emergence of Cooperation

Since both repeated cooperative behavior and repeated betrayal behavior are equilibria, what distinguishes them?

It is *trust*. In the cooperative case both players trust each other, while in the betrayal case they do not.

So, the *emergence of cooperation* comes down to the *emergence of trust*. Cooperation is impossible without mutual trust.

“Agent A trusts agent B”. Can this statement be explained further? It means A *expects* B to behave well in future encounters.

So trust can be regarded as an *expectation* of behavior. It is therefore no surprise that probability is useful in modeling social agents.

## Trust and Experience

How does trust emerge? From *repeated good experience*. Agent A's trust in agent B increases (decreases) each time B behaves well (badly) in a mutual encounter.

So it is a *sequence* of experiences that are encoded into *trust* somehow. The mathematical theory of how that is done is on-going research, but some requirements are widely accepted.

**Example** Say agent A has interacted with both agents B and C, with the same number of interactions. Suppose A's experience sequence of B is at least as good as of his with C at each point in the sequence. Then A's trust in B is at least as good as A's trust in C.

## Trust Dynamics

Agents' trust of each other can – and usually will – change as they interact through time. This much is obvious. Moreover, if B behaves badly to A in transactions, A's trust in B will reduce. Conversely A's trust in B will increase if B keeps behaving well.

How trust changes as the interactions proceed is a topic of much current research. The area is called *Trust Dynamics*.

There is another way to think of A's trust in B – “A *believes* B will behave well in transactions with A.” Thus trust can also be regarded as a *belief*.

Therefore when trust changes it is a form of *Belief Revision*, which is a very active area in AI.

## Trust Update Functions

All models of trust dynamics have this basic structure. Let the *trust space* be  $T$ . Let *sequences of experience judgements* be in the space  $ES$ . Then a *trust update function*  $F$  is a map:

$$F : T \times ES \rightarrow T.$$

$F$  takes as inputs the last trust value together with the the entire sequence of experience judgments up to the current one and uses them to produce a new trust value.

Commonly,  $T$  is modelled by the real intervals  $[-1, +1]$  or  $[0, 1]$ , or by discrete integer values in some range.

## History Matters

Why do we need the entire history of experiences to update trust? Because *memory* is important.

Suppose there are two histories

$$h = \langle e_1, e_2, \dots, e_n \rangle \text{ and } h' = \langle e'_1, e'_2, \dots, e'_n \rangle.$$

These may be the histories of A's interactions with B and C. How should one compare  $h$  (A's judgment of B over time) with  $h'$  (A's judgment of C over time).

I will ask you at this point for your intuitions.

Here are some obvious questions.

If at the beginning B's history is good but over time it's history gets worse and worse, while for C it is the opposite, what should A do about it's respective trust of them? History matters!

How about an erratic agent?

How much "memory" is important? *Discount factor* idea from economics or accountancy.

## Reputation

“He has a good reputation.” “The Minister’s reputation is in decline.” “Westpac has a reputation to protect.”

Intuitively these mean degrees of *trustworthiness*. So, while *trust* and *reputation* are related, are they the same?

Trust starts from *pairwise* encounters. But it is possible for A to trust B even though B has a bad reputation. How so? Think of honor among thieves or the Mafia. The Mafia has a bad reputation because most people do not trust it, but within itself mutual trust was what made it hard to defeat. This is the clue to the difference between trust and reputation.

Reputation is a *community* assessment. The interplay between trust and reputation is complex.

Reputation has its own dynamics. Here is one way to do it. Suppose every agent in that community were to provide trust values for all other agents. From that, reputation scores are computed for every agent. How should that be done? What, if anything, should be done about the trust values provided by agents which have very low reputation? How many times do we iterate?

## Conclusion

In the time available and to keep this talk accessible to everyone I have omitted many relevant topics and details.

I hope to cover them in the tutorial on Social Agents.

Here are some places where you can find more information. Conferences on web trust and technologies, typically sponsored by the IEEE. Conferences on multiagent systems. Conferences on knowledge representation and reasoning. The major AI conferences – IJCAI, AAAI, ECAI, KR. The regional conference PRICAI. Various IEEE Transactions.