Indirect Reciprocity and Costly Assessment in Multiagent Systems

Fernando P. Santos[1,2] and **Jorge M. Pacheco**[3] and **Francisco C. Santos**[1]
[1] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, 2744-016 Porto Salvo, Portugal
[2] **Email:** fernando.pedro@tecnico.ulisboa.pt, **Phone Number:** (+351) 210 407 091
URL: http://web.ist.utl.pt/fernando.pedro/
[3] CBMA and Departamento de Matemática e Aplicações, Universidade do Minho, 4710-057 Braga, Portugal

Abstract

Social norms can help solving cooperation dilemmas, constituting a key ingredient in systems of indirect reciprocity (IR). Under IR, agents are associated with different reputations, whose attribution depends on socially adopted norms that judge behaviors as *good* or *bad*. While the pros and cons of having a certain public image depend on how agents learn to discriminate between reputations, the mechanisms incentivizing agents to report the outcome of their interactions remain unclear, especially when reporting involves a cost (costly reputation building). Here we develop a new model – inspired in evolutionary game theory – and show that two social norms can sustain high levels of cooperation, even if reputation building is costly. For that, agents must be able to anticipate the reporting intentions of their opponents. Cooperation depends sensitively on both the cost of reporting and the accuracy level of reporting anticipation.

Introduction

Social norms are a cornerstone of human societies, being a fundamental mechanism to solve coordination, cooperation and collective action problems. In general, social norms are public and establish an expected pattern of behavior. When violated, they may lead to responses that range from gossip to open censure, ostracism, or dishonor for the transgressor (Bicchieri 2005). Often, norms that prevail in a society only enforce behaviors indirectly, functioning as a top-down mechanism that influences the bottom-up adherence (or not) to certain behaviors. This is particularly evident when systems of reputations are used to enforce social norms (Castelfranchi, Conte, and Paolucci 1998): acting in a certain way may provide a reputation uplift/downgrade whose tangible effect emerges in the future, as a form of reciprocation. In fact, enforcing behaviors indirectly, through norms and reputations, underlies Indirect Reciprocity (IR), known as a fundamental mechanism for the evolution of cooperation among humans (Nowak and Sigmund 2005).

IR shares a fundamental challenge with other reputation systems working as enforcement mechanisms of social norms: they require observability (Haynes et al. 2017), *i.e.*, agents able to observe the behaviors of their peers. With rare exceptions (Sasaki, Okada, and Nakai 2016), previous models typically assume that observability is an exogenous factor. In reality, however, accessing the information about a private interaction depends on the decision of the involved agents that may share (or not) its outcome. As an example, in *e-commerce* or *p2p* platforms, private interactions take place and the individuals need to be incentivized to rate their opponents, that is, to provide information about their actions. This naturally involves time and effort. When the process of information sharing is costly, reporting is hardly fulfilled by rational agents, such that the system of IR – and ensuing cooperation – may collapse. Here we address the problem of costly reputation building in multiagent systems and the sustainability of cooperation. We show that, even with costly reporting, cooperation can emerge, provided that agents are able to anticipate the reporting intentions of their opponents.

Model Summary

We consider a population of Z agents that play (in pairs, selected uniformly at random) a donation game in which they must Cooperate (C) or Defect (D) with each other. Cooperation involves paying a cost (c) in order to provide a benefit (b) to the opponent ($b > c > 0$), whereas Defection does not imply any cost/benefit. The social optimum is achieved when both agents cooperate; yet, for each agent, the fact that cooperation is costly configures defection as a preferred option. This originates a social dilemma (Prisoner's Dilemma). On top of that, agents have a reputation ($Good$, G or Bad, B). After each interaction, the reputation of an agent X, who decided C or D against another agent Y, will change given a social norm, that is, a rule used to attribute a new reputation to X, given the actions and characteristics of X and Y. The reputation update will only occur provided that Y shared the outcome of the interaction (at a cost of reporting $c_R > 0$). Additionally, we provide agents the possibility to anticipate the intention of their opponents in sharing information about the interaction. While anticipative decision-making often requires complex anticipation architectures that allow predicting future outcomes based on the past, in the present domain (*e.g.*, *e-commerce* platforms and, in general, any artificial agent society with reputation systems) this can be achieved by making publicly available the previous reports of agents.

Strategies (i.e., rules that dictate whether to C or D against a G or B opponent) evolve following a process of social

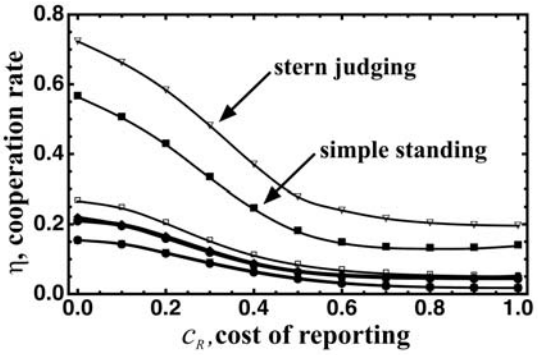

Figure 1: Cooperation emerges even with costly assessment, particularly with two social norms. $b = 5$, $c = 1$, $Z = 50$, $\tau = 0.01$.

learning: from time to time, pairs of agents (A and B) are randomly sampled from the population and A adopts the strategy of B with a probability that grows with fitness difference ($f_B - f_A$). Successful strategies spread faster in the population. Very rarely, agents explore the strategy space, without resorting to fitness comparison (Santos, Pacheco, and Santos 2016). We keep track of the stationary cooperation rate, i.e., the average level of cooperation once the system reaches a steady state, provided the stationary distribution of reputations (Santos, Santos, and Pacheco 2016).

Results

We show that the mechanism of *reporting anticipation* suffices to sustain cooperation under IR and costly reputation building ($c_R > 0$). There are social norms that efficiently allow cooperation to be sustained (Santos, Pacheco, and Santos 2018). In particular, the so-called *stern judging* (an agent is G if Cooperated with G or Defected with B; all else is B) is the social norm allowing the highest values of cooperation, regardless of the anticipation error τ and the reporting cost c_R. A more benevolent norm often called *simple standing*, that regards as G also those that cooperated with B opponents, is the second norm promoting the highest levels of cooperation. Finally, we show that cooperation under IR and costly reputation building is fundamentally sensitive to the cost of reporting (c_R) and anticipation errors (τ). In a previous work, these two norms were found to promote high levels of cooperation in finite populations (Santos, Santos, and Pacheco 2016) and for a large range of exploration rates by agents (Santos, Pacheco, and Santos 2016).

Conclusion and Future Work

Here we investigate whether indirect reciprocity can promote cooperation when reputation building is costly. We develop an evolutionary game theoretical model which describes the dynamics of strategy adoption, when agents' reputations are governed by different social norms. Importantly, this new model allows us to understand which social norms promote cooperation, and why. We conclude that cooperation can emerge with indirect reciprocity, even if reputation building is costly, provided that agents are able to anticipate the reporting intentions of their opponents.

In methodological terms, we contribute with a novel analytical framework that allows studying the interplay between social norms and cooperation, while avoiding the burden of large-scale simulations. The framework models an environment in which interaction observability is costly and depends on agents' decision, opening the opportunity for studying central aspects of social norms, reputation systems and cooperation. Future extensions may include, for example, the role of social norms that prevent/instigate malicious reports and lying, new incentives for honest reporting, or even take into account the role of *bluffing*, when signalling the intention of reporting an action does not translate in an actual report. The presented framework can also be applied to other dilemmas, such as coordination, co-existence or public good games. Finally, while here we assume a direct link between the reporting intention of a Receiver and the ability of a Donor to anticipate that purpose, we shall extend this framework in future work, e.g., to accommodate more complex anticipation architectures (Domingos, Burguillo, and Lenaerts 2017).

Acknowledgments

This research was supported by FCT-Portugal through grants SFRH/BD/94736/2013, PTDC/EEI-SII/5081/2014, PTDC/MAT/STA/3358/2014, UID/BIA/04050/2013, and UID/CEC/50021/2013.

References

Bicchieri, C. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Castelfranchi, C.; Conte, R.; and Paolucci, M. 1998. Normative reputation and the costs of compliance. *J Artif Soc Soc Simulat* 1(3):3.

Domingos, E. F.; Burguillo, J.-C.; and Lenaerts, T. 2017. Reactive versus anticipative decision making in a novel gift-giving game. In *AAAI'17*, 4399–4405. AAAI Press.

Haynes, C.; Luck, M.; McBurney, P.; Mahmoud, S.; Vítek, T.; and Miles, S. 2017. Engineering the emergence of norms: a review. *Knowl Eng Rev* 32.

Nowak, M. A., and Sigmund, K. 2005. Evolution of indirect reciprocity. *Nature* 437(7063):1291–1298.

Santos, F. P.; Pacheco, J. M.; and Santos, F. C. 2016. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Sci Rep* 6(37517).

Santos, F. P.; Pacheco, J. M.; and Santos, F. C. 2018. Social norms of cooperation with costly reputation building. In *AAAI'18*, (to appear). AAAI Press.

Santos, F. P.; Santos, F. C.; and Pacheco, J. M. 2016. Social norms of cooperation in small-scale societies. *PLoS Comput Biol* 12(1):e1004709.

Sasaki, T.; Okada, I.; and Nakai, Y. 2016. Indirect reciprocity can overcome free-rider problems on costly moral assessment. *Biol Lett* 12(7):20160341.