Successful social norms of indirect reciprocity beyond binary reputation

Yohsuke Murase[1], Minjae Kim[2], and Seung Ki Baek[2]

[1]RIKEN Center for Computational Science, Kobe, Hyogo, Japan
yohsuke.murase@gmail.com
[2]Department of Physics, Pukyong National University, Busan, Korea

Abstract

Indirect reciprocity is one of the fundamental mechanisms that promotes cooperative behavior among self-interested individuals by means of reputation when cooperative behaviour benefits the society at an individual cost. Most of the previous studies have assumed that reputation is either good or bad, but such a binary-reputation system is a crude approximation to reality. In this work, we add another reputation called 'neutral' and fully identify ternary norms that achieve cooperation and possess evolutionary stability against behavioural mutants. Comparison with the results from the binary-reputation system suggests universal features of successful norms, that is, maintenance of cooperation by the majority, identification of defectors to punish them, justification of the punishment, and apology with forgiveness.

Introduction

The ability to cooperate with others is a key trait to make our society highly effective, and indirect reciprocity is one of the most fundamental mechanisms to tackle this task. Let us consider an infinitely large population in which the donation game is repeatedly played between two randomly picked players, one as a 'donor' and the other as a 'recipient'. The donor chooses to either cooperate (C) or defect (D): If C is chosen, the donor provides a benefit of b for the recipient at a cost of c ($b > c$). Otherwise, their payoffs are both zero. Now, in the presence of indirect reciprocity, a social norm comes into play: Based on the social norm, a donor decides what to do to the recipient by referring to their own reputations, and an observer updates reputation after observing who did what against whom. A successful social norm provides a strong incentive for actions that are costly but beneficial to a co-player because a well-reputed player has a high chance to be rewarded by others. A central question is what are the requirements for social norms to achieve stable cooperation.

The leading eight are successful social norms that maintain cooperation in the binary-reputation system for any benefit-to-cost ratio b/c (Ohtsuki and Iwasa, 2004, 2006). They have evolutionary stability in the sense that a mutant that does not follow the prescribed behaviour will fail to achieve fixation when one of the leading eight prevails in the society. Their rules are summarized in Table 1.

Table 1: Prescriptions that are commonly shared by the leading eight. The asterisk ($*$) is a wildcard, meaning that it can be any of G and B. The left two columns show reputations, and the third column is the action A prescribed by the behavioural rule. The fourth column indicates the reputation assigned to the donor who executed the action A, whereas the last column shows the reputation resulting from the other action $\neg A$. The dagger ($\dagger$) means that the action is either C or D depending on the assignment rule, so it is C if and only if $R(B,B,C) = G$ and $R(B,B,D) = B$.

X	Y	A	$R(X,Y,A)$	$R(X,Y,\neg A)$
G	G	C	G	B
G	B	D	G	$*$
B	G	C	G	B
B	B	$\dagger$	$*$	$*$

Since the beginning of mathematical analysis of indirect reciprocity (Nowak and Sigmund, 1998), most of the previous studies on indirect reciprocity, including the leading eight, have assumed that reputation is either 'good'(G) or 'bad'(B), and the extension beyond such binarity is relatively rare in the literature (Tanabe et al., 2013; Lee et al., 2021). Although the binary-reputation are conceptually simple, such dichotomy may be a crude approximation of reality if we consider the large grey area between good and bad. Furthermore, it is not always clear how the conclusions from the binary-reputation system generalize because they may be consequences of the oversimplification.

What would be the universal characteristics that every successful norm shares irrespective of the form of reputation? How should we revise the conclusion learned from the binary-reputation system when the binarity assumption is relaxed? To get insights into these questions, we study a ternary-reputation system, in which players are labeled by three types of reputation (Murase et al., 2021). As has been done to find the leading eight, we fully identify evolutionarily stable norms that achieve cooperation through

direct enumeration. Because the strategy space expands super-exponentially as the number of possible reputations increases, the enumeration requires massive computation with a supercomputer. Our result shows both similarity and dissimilarity between binary- and the ternary-reputation systems, suggesting universal features of successful norms as well as limitations of the binary system.

Main results

We have three labels for representing reputation, i.e., G (good), N (neutral), and B (bad). However, note that no ordinal relationships is assumed among them so that N may be worse than B, for instance. A social norm is comprised of an assessment rule and a behavioural rule: An assessment rule determines a donor's new reputation after observing the donor's interaction with the recipient, and the rule is represented by a map $R(X, Y, A) \to Z$, where $X, Y \in \{G, N, B\}$ are reputations of the donor and the recipient, respectively, and $A \in \{C, D\}$ is the donor's action. Likewise, a behavioural rule is represented by a map $P(X, Y) \to A$, where X and Y are reputations of the donor and the recipient, respectively, and A is the prescribed action. Because of 18 possible combination of (X, Y, A), we have $3^{18} = 387,420,489$ assignment rules, and 2^9 possible behavioural rules exist for each assignment rule. Thus, the total number of social norms amounts to $3^{18} \times 2^9 = 198,359,290,368$.

Among these social norms, we comprehensively identified the ones that achieve a sufficiently high level of cooperation as well as evolutionary stability against every mutant having a different behavioural rule. We found roughly 1.8 million social norms that have the above properties for any b/c. To better characterize these norms, we propose a classification scheme based on their cooperation, punishment, and recovery patterns, which divides them into 14 classes. An example is shown in Table 2.

Table 2: An example of successful social norms with the ternary-reputation system of G (good), N (neutral), and B (bad).

X	Y	A	$R(X,Y,A)$	$R(X,Y,\neg A)$
G	G	C	G	B
G	N	C	G	B
G	B	D	N	B
N	G	C	G	B
N	N	D	N	N
N	B	D	N	B
B	G	C	N	B
B	N	D	B	B
B	B	D	N	B

Let us recall that the leading eight have the following common characteristics: (i) Maintenance of cooperation, (ii) Identification of defectors, (iii) Punishment and its justification, and (iv) Apology and forgiveness. Overall, these characteristics are also common in the ternary case, suggesting their universality in a general reputation system. However, it should be noted that some of the above characteristics are relaxed in the ternary system. First, a cooperative population may have more than one type of reputation in equilibrium, whereas most players keep good reputation in case of the leading eight. Second, a partial justification of punishment is also allowed: With the leading eight, a player can maintain good reputation even if he or she defects against an ill-reputed player. On the other hand, according to some of the successful ternary norms, a player who carries out punishment may lose good reputation. Third, it may take more than one time step to recover reputation: It is in contrast with the case of the leading eight, in which an ill-reputed player is forgiven immediately after donating to a well-reputed player because it is the only possible way to forgive the ill-reputed player within the binary system.

In summary, based on the results from the ternary-reputation system, we conjecture that successful norms with a general reputation system should share the following characteristics:

1. Maintenance of cooperation by the majority (but not necessarily all) of the population.
2. Identification of defectors.
3. Punishment and its partial or full justification.
4. Apology and forgiveness, either gradual or instantaneous.

These rules will serve as guiding principles even when we design a social norm based on a reputation system with finer gradations.

It remains as an open question whether the strategies found in this study can achieve a high level of cooperation when reputation is assigned privately rather than publicly (Hilbe et al., 2018), and further research is called for to explore this possibility.

Acknowledgement

Part of the results is obtained by using the Fugaku computer at RIKEN Center for Computational Science (Proposal number ra000002). We appreciate the APCTP for its hospitality during completion of this work. Y.M. acknowledges support from Japan Society for the Promotion of Science (www.jsps.go.jp) (JSPS KAKENHI; Grant no. 18H03621 and 21K03362). S.K.B. acknowledges support by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (www.moe.go.kr) (NRF-2020R1I1A2071670).

References

Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K., and Nowak, M. A. (2018). Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl. Acad. Sci. USA*, 115(48):12241–12246.

Lee, S., Murase, Y., and Baek, S. K. (2021). Local stability of cooperation in a continuous model of indirect reciprocity. *arXiv preprint arXiv:2104.02881*.

Murase, Y., Kim, M., and Baek, S. K. (2021). in preparation.

Nowak, M. A. and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685):573.

Ohtsuki, H. and Iwasa, Y. (2004). How should we define goodness? – reputation dynamics in indirect reciprocity. *J. Theor. Biol.*, 231(1):107–120.

Ohtsuki, H. and Iwasa, Y. (2006). The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.*, 239(4):435–444.

Tanabe, S., Suzuki, H., and Masuda, N. (2013). Indirect reciprocity with trinary reputations. *J. Theor. Biol.*, 317:338–347.