Social Network Interventions to Prevent Reciprocity-driven Polarization

Extended Abstract

Fernando P. Santos
Department of Ecology and Evolutionary Biology
Princeton University
Informatics Institute, University of Amsterdam
fpsantos@princeton.edu

Francisco C. Santos
INESC-ID & Instituto Superior Técnico
Universidade de Lisboa
franciscocsantos@tecnico.ulisboa.pt

Jorge M. Pacheco
CBMA & Department of Mathematics
University of Minho
jmpacheco@math.uminho.pt

Simon A. Levin
Department of Ecology and Evolutionary Biology
Princeton University
slevin@princeton.edu

ABSTRACT

Complex networks and reputation systems are fundamental mechanisms to sustain cooperation in populations of self-regarding agents. These mechanisms are typically studied in isolation. In online social platforms, however, behavioral dynamics are likely to result from their combination. Here we investigate the relationship between social networks and reputation-based cooperation (in a Prisoner's Dilemma setting) in large populations. We develop a new evolutionary game-theoretical model and study dynamics in networks with varying degrees of community structure. We show that networks exhibiting modular structures hamper global cooperation: reputation-based group identities emerge in different communities and strategies that uniquely cooperate with in-group members fixate, sustaining polarization and group bias. Global cooperation is recovered provided that inter-community edges are added.

KEYWORDS

Cooperation; Evolutionary Game Theory; Complex networks; Polarization; Reputations; Social norms; Indirect reciprocity

ACM Reference Format:
Fernando P. Santos, Francisco C. Santos, Jorge M. Pacheco, and Simon A. Levin. 2021. Social Network Interventions to Prevent Reciprocity-driven Polarization: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 INTRODUCTION

A significant fraction of social interactions occur, nowadays, in online platforms, along the edges of social networks and often mediated by reputation systems [11]. Networks [6, 12] and reputations [5] are fundamental mechanisms for cooperation [10]. When information about individuals' past behaviors (reputation) is available, cooperation can be sustained through indirect reciprocity [5]. In this context, reputations are attributed based on social norms that work as rules defining which actions — and against whom — lead to positive/negative reputations [7, 14]. The behavioral dynamics resulting from combining networks and indirect reciprocity are not trivial and remain scarcely explored — for some exceptions see [2, 4, 9]. Here, we analyze the impact of social network topology and indirect reciprocity on cooperation and polarization.

We generate graphs that span different classes of community structure, from topologies with well-defined modules to ones with loosely defined communities. Over such networks, we simulate evolutionary game dynamics of cooperation and indirect reciprocity. We show that, when communities are well-defined and reputations are attributed following one of the leading social norms in promoting cooperation — called *stern-judging* [8] — polarization and group bias emerge: cooperation thrives within communities, though not across communities. This situation is resolved if connections between individuals belonging to different communities are established. When other social norms are considered (e.g., image-score or simple-standing) cooperation is poorer inside communities, compared to stern-judging, although they require smaller numbers of inter-community links to recover high levels of cooperation.

2 MODEL

We consider a population with Z individuals. Each one is initially attributed (randomly) a reputation in the eyes of others — Good (G) or Bad (B) — and a strategy. This leads to the following four possible strategies: unconditional Defection (*AllD*), unconditional Cooperation (*AllC*), Discriminator strategy (*Disc*), that is, cooperate with those with good reputation, and defect otherwise), and paradoxical Discriminator strategy (*pDisc*, the opposite of *Disc*). These strategies define the actions in a Donation Game, where a Donor and a Recipient interact. When Donors cooperate (C) they will pay a cost c and the Recipients receive a benefit b. If Donors defect (D), both individuals get 0. As we assume that $b > c > 0$, when two individuals play a Donation Game both in the role of Donor and Recipient, we obtain the famous Prisoner's Dilemma.

Individuals interact in a graph $G = (N, E)$. Each individual i corresponds to a node $n_i \in N$. If an edge $e_{ij} \in E$ exists, i may play, imitate, or share reputations with j. We follow a procedure closely related to the generation of Caveman Graphs [23] and the Girvan-Newman benchmark [3]. We generate n homogeneous random

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

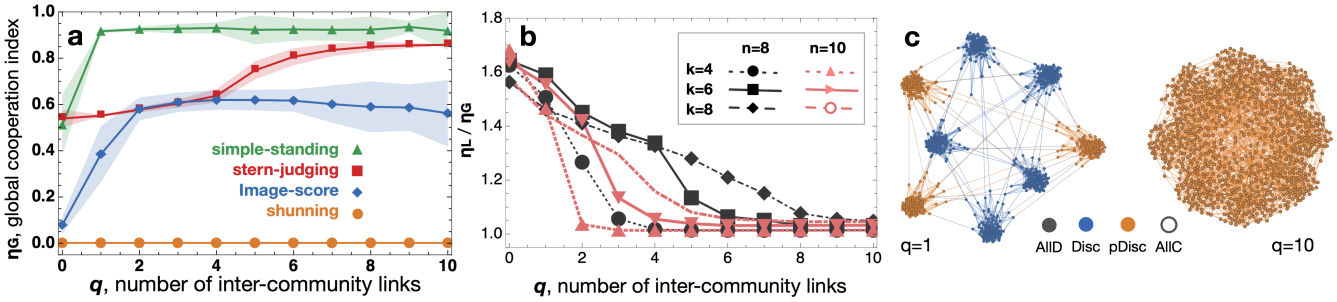

Figure 1: a) Global cooperation, η_G; for low q stern-judging leads to group bias (high η_L and low η_G). As q increases, global cooperation is recovered. b) Local over global cooperation ratio under stern-judging, for different combinations of community count ($n = \{8, 10\}$) and network degree ($k = \{4, 6, 8\}$). c) Example of networks generated with a different q and illustration of evolving strategies, where structural polarization [20] is visible for low $q = 1$. $Z = 1000$, $b = 5$, $c = 1$, $\mu = 0.001$, $n = 8$, $k = 6$.

networks (what we call modules) with an average degree k. For each pair of modules, we then swap the endpoints of q randomly chosen (distinct) pairs of edges, one in each module, forcing the creation of inter-community edges and keeping degree distribution. This way, q and k control the relative fraction of inter (over intra)-community links. Higher q and lower k imply that communities are more loosely defined.

At each time-step, a random individual is selected to revise her strategy by imitating a neighbor [19]. Strategy revision depends on payoff accumulated in Donation Games; neighbors performing better have higher probabilities of being imitated. With probability μ (mutation/exploration) a random strategy is adopted [13].

The reputation of a Donor is updated based on 1) her action, 2) the reputation of the Recipient and 3) the social norms employed by her neighbors. We consider that a social norm stands as a rule that dictates the expected behavior of agents that act as donors, defining how to attribute a new reputation (Good, G or Bad, B), given their action (Cooperate, C or Defect, D) and the reputation of the recipient (G or B) [14]. Social norms encoding this type of information are classified as second-order norms [7, 16]. Four social norms have been given special attention: Stern-judging, which assigns a reputation G to a donor that helps a G recipient or refuses help to a B one; Simple-Standing (SS), that only assigns a reputation B to a Donor that defects with a G Recipient; Shunning (SH), similar to SJ but less "benevolent", by also assigning B to any donor that defects; and Image Score (IS) where all that matters is the action of the Donor, who acquires a reputation G if playing C and a reputation B if playing D. After each interaction between a Donor i and a Recipient j, we select a randomly neighbor of i to be the Observer (o). The Observer judges i following i's action and the opinion that o has over j. This information spreads to the neighbors of o (*e.g.*, gossip) that will update their view of i according to the reputation shared by o. We consider *execution*, *assessment* and *assignment* errors (all with probability 0.01) as described in [15].

3 RESULTS

We observe that, despite supporting a high fraction of cooperative acts, stern-judging leads to socially polarized situations where cooperation is only high in local communities. To measure this, we compute a global index of cooperation ($\eta_G, 0 \leq \eta_G \leq 1$). For a given individual i, we use Y_i as the number of individuals in the whole population that would cooperate with i and N_i ($N_i = Z - 1 - Y_i$) as the number of defectors against i. We define $\eta_G = (Z-1)^{-1} \sum_{i=1}^{Z-1} \frac{Y_i}{Z-1}$. η_L, local cooperation, is the fraction of cooperative actions along edges. By comparing η_L and η_G, we can quantify the level of group bias in the population: High η_L and low η_G means that cooperation tends to be only local. Figure 1 shows that, for low values of q, stern-judging leads to high η_L and low η_G.

4 CONCLUSION

Here we show that reputation dynamics on top of particular social networks can be detrimental for global cooperation by inducing social polarization. Using synthetic networks, we show that graphs revealing well-defined communities are a suitable environment for individuals to adopt antagonistic strategies and develop group bias. Under stern-judging, individuals end up condemning cooperative acts with the out-group (*i.e.*, outside their close community). The way strategies evolve to condition cooperation based on G and B results from a convention [17, 18, 21, 22]. The model that we present can be extended to accommodate more complex mechanisms and dynamics. It would be interesting to study multiple norms co-existing in the population, particularly in different communities. This suggests the implementation of multi-level evolutionary dynamic models [1, 8, 24], where norms and strategies co-evolve.

All in all, we hope that this model may broaden our knowledge on the effective design of multiagent systems where global cooperation emerges, and of link-rewiring algorithms that contribute to sustain long-term pro-sociality and to avert out-group conflict.

ACKNOWLEDGMENTS

This work was supported by FCT-Portugal (grants nos. UIDB/50021/2020, PTDC/MAT-APL/6804/2020, and PTDC/CCI-INF/7366/2020), by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Dynamic and Multi-scale Systems Postdoctoral Fellowship Award and Collaborative Award, the National Science Foundation (grant no. CCF1917819), the C3.ai Inc. and Microsoft Corporation and the Army Research Office (grant no. W911NF-18-1-0325).

REFERENCES

[1] Daniel B Cooney. 2019. The replicator dynamics for multilevel selection in evolutionary games. *J. Math. Biol.* 79, 1 (2019), 101–154.

[2] Feng Fu, Christoph Hauert, Martin A Nowak, and Long Wang. 2008. Reputation-based partner choice promotes cooperation in social networks. *Phys. Rev. E* 78, 2 (2008), 026117.

[3] Michelle Girvan and Mark EJ Newman. 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99, 12 (2002).

[4] Jörg Gross and Carsten KW De Dreu. 2019. The rise and fall of cooperation through reputation and group polarization. *Nat. Commun* 10, 1 (2019), 776.

[5] Martin A Nowak and Karl Sigmund. 2005. Evolution of indirect reciprocity. *Nature* 437, 7063 (2005), 1291.

[6] Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A Nowak. 2006. A simple rule for the evolution of cooperation on graphs and social networks. *Nature* 441, 7092 (2006), 502–505.

[7] Hisashi Ohtsuki and Yoh Iwasa. 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. *J. Theor. Biol.* 231, 1 (2004), 107–120.

[8] Jorge M Pacheco, Francisco C Santos, and Fabio AC C Chalub. 2006. Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLOS Comput. Biol.* 2, 12 (2006), e178.

[9] Ana Peleteiro, Juan C Burguillo, and Siang Yew Chong. 2014. Exploring indirect reciprocity in complex networks using coalitions and rewiring. In *Proc of AAMAS'14*. International Foundation for Autonomous Agents and Multiagent Systems, 669–676.

[10] David G Rand and Martin A Nowak. 2013. Human cooperation. *Trends in cognitive sciences* 17, 8 (2013), 413–425.

[11] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48.

[12] Francisco C Santos and Jorge M Pacheco. 2005. Scale-free networks provide a unifying framework for the emergence of cooperation. *Phys. Rev. Lett.* 95, 9 (2005), 098104.

[13] Fernando P Santos, Jorge M Pacheco, and Francisco C Santos. 2016. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Sci. Rep.* 6 (2016), 37517.

[14] Fernando P Santos, Jorge M Pacheco, and Francisco C Santos. 2018. Social norms of cooperation with costly reputation building. In *AAAI'18*.

[15] Fernando P Santos, Francisco C Santos, and Jorge M Pacheco. 2016. Social norms of cooperation in small-scale societies. *PLoS Comput. Biol.* 12, 1 (2016), e1004709.

[16] Fernando P Santos, Francisco C Santos, and Jorge M Pacheco. 2018. Social norm complexity and past reputations in the evolution of cooperation. *Nature* 555, 7695 (2018), 242.

[17] Sandip Sen and Stéphane Airiau. 2007. Emergence of norms through social learning. In *IJCAI*, Vol. 1507. 1512.

[18] Brian Skyrms. 2014. *Evolution of the social contract.* Cambridge University Press.

[19] Arne Traulsen, Martin A Nowak, and Jorge M Pacheco. 2006. Stochastic dynamics of invasion and fixation. *Phys. Rev. E* 74, 1 (2006), 011909.

[20] Vítor V Vasconcelos, Simon A Levin, and Flávio L Pinheiro. 2019. Consensus and polarization in competing complex contagion processes. *J. R. Soc. Interface* 16, 155 (2019).

[21] Daniel Villatoro, Jordi Sabater-Mir, and Sandip Sen. 2011. Social instruments for robust convention emergence. In *Proc of IJCAI'11*.

[22] Yixi Wang, Wenhuan Lu, Jianye Hao, Jianguo Wei, and Ho-fung Leung. 2018. Efficient convention emergence through decoupled reinforcement social learning with teacher-student mechanism. In *Proc. of AAMAS'18*. International Foundation for Autonomous Agents and Multiagent Systems, 795–803.

[23] Duncan J Watts. 1999. Networks, dynamics, and the small-world phenomenon. *Am. J. Sociol.* 105, 2 (1999), 493.

[24] Jason Xu, Julian García, and Toby Handfield. 2019. Cooperation with Bottom-up Reputation Dynamics. In *Proceedings of AAMAS'19*. International Foundation for Autonomous Agents and Multiagent Systems, 269–276.