Attraction and Cooperation in Noisy Environments: Individuals and Groups

Elpida Tzafestas

Institute of Communication and Computer Systems
National Technical University of Athens
Zographou Campus, Athens 15773, GREECE
brensham@softlab.ece.ntua.gr

Abstract. This paper presents an attraction mechanism and its impact on cooperation in a society of agents. We adopt the benchmark setting of the Iterated Prisoner's Dilemma (IPD, [1]) and we implement attraction as a mechanism that changes the usual agent's behavior (strategy). More specifically, an agent follows its regular strategy unless it faces an attractive agent. In the latter case, she becomes unconditionally cooperative. This mechanism is shown to yield higher average agent scores in tournaments within uniform or mixed populations than if it were not present. Benefits are higher for higher attraction factors, bigger populations or populations of "irrational" agents. We have also studied the impact of the attraction pairing type and we have found that random not necessarily reciprocal pairing is the best, because even in the absence of reciprocity a rational agent can cooperate consistently with a cooperating attracted, even if irrational, opponent. We have experimented with extreme types of agents as well, namely, a "Don Juan" agent that is attracted by all others, and a "Sex Symbol" to whom all others are attracted. Very often, the introduction of a single extreme agent in a society with no other attraction relations may lead to the same result as a regular society with attraction (but without the extreme agent). Finally, we have experimented with groups of interacting agents to identify possible social conditions that enhance cooperation. All our results suggest that psychological mechanisms external to the economic setting can interfere with it and can actually lead to enhanced cooperation and social stability, despite apparent inconsistencies or irrationalities in individual agents behavior. Accordingly, we can sometimes manage some economic systems by manipulating an external social condition that interferes with normal economic behavior.

Keywords: Cooperation, Iterated Prisoner's Dilemma (IPD), Rationality, Attraction, Don Juan, Sex Symbol, Groups.

1 Introduction

A major research theme in both artificial economics and theoretical biology is the emergence and domination of cooperative behavior between selfish agents. The cooperation problem states that each agent has a strong personal incentive to defect,

while the joint best behavior would be to cooperate. This problem is traditionally modeled as a special two-party game, the Iterated Prisoner's Dilemma (IPD).

At each cycle of a long interaction process, the agents play the Prisoner's Dilemma. Each of the two may either cooperate (C) or defect (D) and is assigned a payoff defined as follows.

Agent	Opponent	Payoff
C	C	3 (= Reward)
C	D	0 (= Sucker)
D	C	5 (= Temptation)
D	D	1 (= Punishment)

The first notable behavior for the IPD designed and studied by Axelrod [1][2] is the Tit For Tat behavior (TFT, in short) :

Start by cooperating,
From there on return the opponent's previous move

This behavior has achieved the highest scores in early tournaments and has been found to be fairly stable in ecological settings. TFT demonstrates three important properties, shared by most high scoring behaviors in IPD experiments.

- *It is good (it starts by cooperating)*
- *It is retaliating (it returns the opponent's defection)*
- *It is generous (it forgets the past if the defecting opponent cooperates again).*

Further strategies include stochastic ones ([9]), the Pavlov strategy ([11]) that cooperates when it has played the same move as its opponent etc. In the literature we may also find studies in an evolutionary perspective ([5]), theoretical or applied biological studies ([3][4][7]) and studies of modified IPD versions ([12]).

We adopt the noisy version of IPD in which there is a nonzero probability that an agent's action will be switched to the opposite, i.e. from *COOPERATE* to *DEFECT* or vice versa. It has been shown that retaliating strategies such as TFT can score quite badly in the presence of noise, despite their superiority in the non-noisy domain [6][8]. This happens because even accidental defections may lead to a persistent series of mutual defections by both players, thus breaking cooperation. The usual approach is to introduce some degree of explicit generosity to account for opponent's misbehaviors or to attempt opponent modeling.

Our approach purports to show that an independent psychological or social factor can allow agents in a society to cooperate fairly well despite noise and without explicit opponent modeling or other intricate reasoning behavior. The organization of the paper is as follows: in section 2 we introduce the attraction mechanism and we discuss the motivation behind it and its conceptual implications. In sections 3 to 5 we describe experiments done with various pairing types, personality extremes and groupings. We finally sum up and discuss our findings in the last section.

2 Rationality and cooperation

Our motivation behind the introduction of an attraction mechanism is the general observation that in human societies, and especially in social contexts, the agents' behavior can be heavily influenced by external psychological and social factors and also many times it can be driven to behaviors outside their normal scope. By "external" we mean a factor or process that is not influenced itself by the primary agent task and does not normally participate in it. We are using the benchmark iterated prisoner's dilemma (IPD) in its noisy version as a study vehicle with a stronger bias toward defection, where we feel it could make sense to introduce such an external attraction factor. More specifically, we believe that biological evolution or, equivalently, social experience would spontaneously exploit any external factor that would induce better agent scores. This is particularly true for noisy environments where agent scores may degrade abruptly, and especially when interactions are lengthier.

The attraction mechanism relies on our everyday experience that people tend to be good and cooperative with other people that attract them and tend to be "regular" with the rest. This translates in our model as:

If (attracted by the opponent) then play ALLC (always cooperate),
Else play as usually (for example, TFT)

We should note that noise is applied to the outcome of this behavior as well. We performed tournament experiments with populations of agents playing a noisy IPD. The agents are interconnected via a "web of attraction" where each agent is connected to (attracted by) a number of others. The normal behavior of an agent is usually one of ALLC, ALLD (always defect), TFT and Adaptive TFT [13], but we have also experimented occasionally with STFT (Suspicious TFT) or other strategies. We experimented with both uniform or mixed populations, whose agents have the same or diverse normal strategies. The reason we use mostly ALLC, ALLD, TFT and Adaptive TFT is that we want to make sure we explore the limits of our attraction mechanism by studying its effect on the extreme behaviors (ALLC and ALLD that act without feedback) as well as on the most intelligent ones (TFT that retaliates immediately and Adaptive TFT that tries to make sense of a situation).

In previous work [13], we have classified usual IPD strategies in two categories: "retaliating" (or "rational") and "irrational". Retaliating strategies are those that basically seek cooperation, but may start by exploring the opponent's reaction to a few initial D moves. For example, the suspicious tit-for-tat (or STFT) strategy starts by defecting, and then plays usual tit-for-tat. On the contrary, irrational strategies are those that do not employ any feedback from the game and play blindly using some innate law. For example, periodic strategies repeat patterns of C's and D's, such as CDD, CCD, CDCD etc. In this sense, ALLC and ALLD are irrational strategies whereas TFT and Adaptive TFT are retaliating ones. We have shown that the Adaptive TFT strategy manages to differentiate between retaliating and irrational strategies and especially between a retaliating strategy and an irrational one that has initially the same behavior. For example it manages to converge to total defection against CDCD, that resembles STFT in the beginning. This feature allows it to achieve much higher scores than TFT in usual diverse social environments. As in real

life, we would expect irrational strategies to profit more from our attraction mechanism or other similar mechanisms and rational ones to be less dependent on such add-ons. In sum, we expect cooperation to be able to emerge in social interactions even in the absence of rationality and good reason.

Before proceeding to describe the experimental setup and the results obtained, we should stress the fact that the attraction mechanism described is in our own terms irrational in that it does not depend on any real feedback of the agent. Our results then suggest that the coupling of reasoning mechanisms with reactive ones (such as attraction, be it physical, emotional, social or other) may be advantageous to social behavior and this is in line with current trends in cognitive and social science.

3 Pairing experiments

The first question is whether reciprocity counts in such webs of attraction. We have thus experimented with three different pairing types:

- **Average pairing**: On average, each agent is attracted by M others and relations are reciprocal
- **Exact pairing**: Each agent is attracted by exactly M others and relations are reciprocal
- **Random pairing**: Each agent is attracted by exactly M others randomly, so not all relations are reciprocal

We characterize our experiments by the number of agents N participating in a tournament, the attraction factor M (number of agents that an agent is attracted to) and the normal strategy of the agents. All tournaments involve two games per each pair of agents including two games of each agent with itself. The length of the noisy IPD game has been set to 100 cycles and the degree of noise to 10%. The results with uniform populations comprising 20 or 40 only ALLD or only TFT or only AdaptiveTFT agents are given below. The corresponding results without attraction or without noise are given for reference. Values stand for average total scores in the society.

STRATEGY, NxM	ATTRACTION				NO NOISE
	RANDOM	AVERAGE	EXACT	NONE	
ALLD, 20x5	6936.7	6360.35	6374.0	4592.95	4000
ALLD, 20x10	9061.15	8072.65	8143.65		
ALLD, 40x5	11658.1	10937.675	10982.85	9181.35	8000
ALLD, 40x10	14002.4	12748.85	12768.55		
ALLD, 40x15	16051.1	14510.475	14591.4		
ALLD, 40x20	17926.725	16286.9	16332.7		
TFT, 20x5	10326.8	10043.35	9984.1	9367.95	12000
TFT, 20x10	11046.35	10563.55	10613.4		
TFT, 40x5	19785.574	19222.6	19248.025	18646.1	24000
TFT, 40x10	20723.15	19820.45	19892.625		
TFT, 40x15	21496.574	20401.45	20501.824		

TFT, 40x20	22176.5	21080.8	21104.375		
Adaptive, 20x5	11029.0	10721.7	10490.85	10249.05	12000
Adaptive, 20x10	11465.0	11036.15	10958.1		
Adaptive, 40x5	21244.35	20713.2	20933.5	20529.55	24000
Adaptive, 40x10	21888.676	21152.725	21253.275		
Adaptive, 40x15	22292.225	21516.426	21678.25		
Adaptive, 40x20	22708.625	22034.426	21968.3		

Our results imply the following:

- There is a clear superiority of the random pairing compared with the two other types of pairing that appear to not differ. This is at first counter-intuitive because we would expect exact pairing to be the best. However, the random pairing induces persistent cooperative behavior against non-reciprocally attracted agents, thus inviting them to cooperation by retaliating. This is why score improvement in the case of TFT is much more pronounced than in the case of Adaptive TFT which is by default more robust to noise.
- We have not presented the results of uniform ALLC societies, because as expected there nothing changes with attraction: the agents are ALLC anyway, the only defections being accidental as implemented via the noise process.
- The ALLD behavior profits from the introduction of attraction because it is given the chance to exploit others. On average an ALLD strategy exploits and is being exploited in a web of attraction so that its average score rises beyond its theoretical equilibrium. This is why the scores with noise improve for the ALLD, unlike what happens with retaliating agents.
- The bigger the attraction factor, the bigger the score improvement for any strategy. Again, the improvement is inversely proportional to the degree of "rationality" of the strategy: an ALLD gains more than a TFT that in turn gains more than Adaptive TFT.

We further experimented with mixes of agent strategies to see to what degree these boundary results are reproduced in random social environments. We are giving below the corresponding results that support our previous conclusions.

STRATEGY MIX*, NxM	**ATTRACTION**				**NO NOISE**
	RANDOM	**AVERAGE**	**EXACT**	**NONE**	
(5,5,5,5)x3	9881.1	9732.9	9613.5	8588.45	9265.0
(5,5,5,5)x5	10209.0	9911.4	9911.7		
(5,5,5,5)x10	11063.6	10756.0	10676.9		
(10,10,10,10)x5	19459.15	19166.55	19300.35	17332.824	18530.0
(10,10,10,10)x10	20724.7	19748.85	19892.8		
(10,10,10,10)x15	21480.5	20722.1	20588.45		
(10,10,10,10)x20	22177.2	21343.65	21146.35		

* (n_1,n_2,n_3,n_4) = (no. of ALLC's, no. of ALLD's, no. of TFT's, no. of ADAPTIVE's)

Finally, it was found that, in mixed populations, the impact of the attraction web can be significant on the resulting score ranking, so that an agent can drop from top to

bottom or climb from bottom to top on reinitialization of the attraction relations. Note also that, in mixed populations with attraction the ranking of an agent is not a function of its degree of "rationality": the cleverest agents do not score best. Rather, an "irrational" agent may be the most performant agent in an appropriate context, whereas an otherwise intelligent agent can sink to the end of the score queue. So, attraction is a mechanism that perturbs the regular social/economic relations and calls for elaborate partner selection to both exploit its opportunities and smoothen its drawbacks.

4 Personality experiments

We subsequently proceeded to define "extreme" agent personalities as far as attraction is concerned. More specifically we defined two new types of agents, the *"Don Juan"* and the *"Sex Symbol"* agents. The first one is attracted by every other agent in the system just like a real womanizer, while the second one is an object of attraction for all other agents, hence it behaves as a sex symbol. We give below some results of the introduction of one Don Juan or one Sex Symbol agent in a society of agents with attraction and with random pairing as has been suggested by the results of the previous section.

STRATEGY, NxM	**ATTRACTION**				**NO NOISE**
	REGULAR	***DON JUAN**	***SEX SYMBOL**	**NONE**	
ALLD, 40x10	14002.4	14348.75 (9917)	14014.25 (34080)	9181.35	8000
ALLD, 40x20	17926.725	18156.2 (14568)	17894.95 (30494)		
TFT, 40x10	20723.15	20807.025 (19890)	20725.975 (24165)	18646.1	24000
TFT, 40x20	22176.5	22249.926 (21098)	22223.176 (23979)		
Adaptive, 40x10	21888.676	22044.875 (21143)	21834.05 (23740)	20529.55	24000
Adaptive, 40x20	22708.625	22934.875 (22250)	22889.975 (23669)		

* *In parenthesis we give the score of the Don Juan or Sex Symbol agent.*

We have also experimented with Don Juan and Sex Symbol agents in a society of agents without attraction in the same experimental setting as above. The results are given in the next page. In sum, our results imply the following.

- In presence of attraction, the introduction of one Don Juan or one Sex Symbol agent has almost no or very little effect. Obviously, the effect further dims with the size of the society or with the attraction factor. In absence of attraction, the introduction of one Don Juan or one Sex Symbol agent has

some effect but this is attenuated and vanishes for bigger society sizes and more performant/rational agents by default.

- The Don Juan agent has consistently the lowest score, while the Sex Symbol agent has consistently the highest one. This is because the Don Juan is easily exploited while the Sex Symbol is a dedicated exploiter. This skews a little the results in the ALLD case, because the actual average score of the rest of the society (without the Don Juan or Sex Symbol agent, respectively) does not improve or even worsens.
- The Don Juan profile can lead an agent to become extremely exploited by fellow agents in particular social settings, and thus obtain very low scores to the benefit of other agents in the system. Such an agent may easily serve as a scapegoat.
- The combination of an ALLD behavior with a Sex Symbol profile can lead an agent to obtain extremely high scores in particular social settings, but this will be to the detriment of other agents in the system. Thus such an agent will appear as cruel.

STRATEGY, N	**ATTRACTION**				**NO NOISE**
	REGULAR (x10)	***ONLY DON JUAN**	***ONLY SEX SYMBOL**	**NONE**	
ALLD, 20	9061.15	5099.85 (1354)	5121.2 (18518)	4592.95	4000
ALLD, 40	14002.4	9694.5 (2144)	9704.5 (37042)	9181,35	8000
TFT, 20	11046.35	9536.85 (11229)	9596.6 (12125)	9367.95	12000
TFT, 40	20723.15	18901.2 (22528)	18902.6 (24228)	18646.1	24000
Adaptive, 20	11465.0	10403.2 (11571)	10308.95 (11856)	10249.05	12000
Adaptive, 40	21888.676	20429.275 (23215)	20554.8 (23846)	20529.55	24000

* *In parenthesis we give the score of the Don Juan or Sex Symbol agent.*

- When more than one Don Juan agents and more than one Sex Symbol agents are present the results are more intricate. In the case of retaliating agents, the Sex Symbol agents are the topmost ranking with the Don Juan agents following just after. This is due to the low scores obtained on average by other agents when facing the Sex Symbol ones. In the case of an ALLD society, as expected from the previous results, the Sex Symbol agents are the topmost ranking while the Don Juan agents are the bottommost ranking. Furthermore, the impact of extreme agents in a mostly retaliating society is minimal.
- In sum, the introduction of extreme agents can potentially influence average society scores and phenomena but influences enormously local results. It is therefore expected to play an important role in systems where results of local

interactions may spread in the society of agents. Such are systems with any form of learning (where past interactions influence future ones) and systems with a spatial interaction structure. Indeed, the influence of extreme agents in a spatial noisy IPD game has been studied in [15] and it was confirmed that spatial structures differ significantly qualitatively between systems with and without extreme agents.

5 Group experiments

Our final set of experiments concerns the impact of attraction groups within a single society. Inspired at first by the initial male-female connotation of the Don Juan concept, we experimented with two types of groupings. The first is a two-sexes society where each agent of one group (females) can only be attracted by agents of the other group (males) and vice versa. The second is a more general "multi-ethnic" society where each agent can only be attracted by agents of the same group and not by agents of the other group. Results are given below for equal cardinalities of groups within the society.

STRATEGY, NxM	**ATTRACTION**				**NO NOISE**
	RANDOM	**TWO SEXES**	**TWO GROUPS**	**NONE**	
ALLD, 20x5	6936.7	6828.65	6772.35	4592.95	4000
ALLD, 20x10	9061.15	8202.3	8200.55		
ALLD, 40x5	11658.1	11582.925	11576.75	9181.35	8000
ALLD, 40x10	14002.4	13601.2	13568.175		
ALLD, 40x15	16051.1	15177.375	15162.75		
ALLD, 40x20	17926.725	16375.4	16382.0		
TFT, 20x5	10326.8	10156.35	10237.95	9367.95	12000
TFT, 20x10	11046.35	10537.35	10563.35		
TFT, 40x5	19785.574	19650.45	19676.0	18646.1	24000
TFT, 40x10	20723.15	20530.55	20449.824		
TFT, 40x15	21496.574	20933.25	20942.375		
TFT, 40x20	22176.5	21087.676	21104.176		
Adaptive, 20x5	11029.0	10805.55	10560.25	10249.05	12000
Adaptive, 20x10	11465.0	11023.6	11112.7		
Adaptive, 40x5	21244.35	21237.65	21171.15	20529.55	24000
Adaptive, 40x10	21888.676	21672.725	21584.35		
Adaptive, 40x15	22292.225	21760.625	21924.225		
Adaptive, 40x20	22708.625	22014.55	22055.225		

Our results imply the following:

- Both types of groupings are comparable in final score and inferior to the original random pairing model within the society. Inferiority is more pronounced for bigger attraction factors. This suggests that more dispersed and random connections among the members of a tournament is advantageous. In the immediate future we plan to study whether groupings

are advantageous in case of uneven tournaments, where an agent interacts more with members of its own group than with others.

- The situation slightly improves when we introduce noise in the attraction grouping, so that with a small probability the agents are attracted by agents of the same "sex" or of the other group. This is an initial indication that, as said above, grouping can be advantageous if it coincides more or less with the interaction structure within the society. The need for agents to dynamically develop their own interaction group of permanent partners and a way of so doing are studied in [16].

6 Conclusions

The motivation behind our work is the hypothesis that psychological mechanisms external to the economic setting can interfere with it and can actually lead to enhanced cooperation and social stability despite environmental noise and agent irrationality. We have presented an attraction mechanism and have studied its behavior in noisy IPD games. The attraction mechanism is coupled with a regular IPD strategy to produce the final agent's behavior and is inspired by our everyday experience that people tend to be good and cooperative with other people that attract them and tend to be "regular" with the rest. More specifically, an agent follows its regular strategy unless it faces an attractive agent. In the latter case, she becomes unconditionally cooperative. As expected, this mechanism is shown to yield higher average agent scores in tournaments within uniform or mixed populations than if it were not present. Benefits are higher for higher attraction factors, bigger populations or populations of "irrational" agents (that do not retaliate or reason, such as ALLD). We have also studied the impact of the attraction pairing type: reciprocal and exact, reciprocal and statistical or statistical and not necessarily reciprocal. We have found that statistical (random) not necessarily reciprocal pairing is the best, because even in the absence of reciprocity a rational agent can cooperate consistently with a cooperating attracted, even if irrational, opponent. We have also experimented with extreme types of agents, namely the "Don Juan" that is attracted by everyone and the "Sex Symbol" that attracts everyone. Some interesting results have been obtained in extreme cases such as the dominance of Sex Symbol agents and the potential of Don Juan agents to act as scapegoats. In any case, the impact of extreme personalities is greatly attenuated in big societies of rational, retaliating agents, but is expected to be crucial in systems where local interactions play an important role. We also experimented with agent groups that can be initially thought of as either sexes or ethnic groups. In none of the two cases and even in the presence of noise in the selection of partners, did we observe any improvement in average society score, thus suggesting that the group structure is not advantageous when it does not match the agent interaction structure. All our results indicate that bigger and more diverse attraction networks contribute significantly to the maintenance of cooperation in the noisy environment.

Related work includes a study of spatial games [15], in the same spirit as in [10][14], where we have shown that irrational agents can indeed survive an

evolutionary process in presence of attraction, and a study of emergence of social groups via attraction-driven partner selection [16], where we have shown how more intricate cognitive functions can prove beneficial and have a bias to emerge in a system with attraction.

References

1. Axelrod, R., and Hamilton, W.D.: The evolution of cooperation, Science 211 (1981) 1390-96
2. Axelrod, R.: The evolution of cooperation. Basic Books (1984)
3. Axelrod, R., and Dion, D.: The further evolution of cooperation, Science 242 (1988) 1385-90
4. Feldman, M.W., and Thomas, E.A.C.: Behavior-dependent contexts for repeated plays of the prisoner's dilemma II: Dynamical aspects of the evolution of cooperation, Journal of Theoretical Biology 128 (1987) 297-315
5. Fogel, D.: Evolving behaviors in the iterated prisoner's dilemma, Evolutionary Computation 1 (1993) 77-97
6. Kraines, D., and Kraines, V: Evolution of learning among Pavlov strategies in a competitive environment with noise, Journal of Conflict Resolution, 39(3):439-466 (1995)
7. Milinski, M.: Tit for tat in sticklebacks and the evolution of cooperation, Nature 325 (1987) 433-435
8. Molander, P.: The optimal level of generosity in a selfish, uncertain environment, Journal of Conflict Resolution, (31(4):692-724 (1987)
9. Nowak, M.A., and Sigmund, K.: Tit for tat in heterogeneous populations, Nature 355 (1992) 250-53
10. Nowak, M.A., and May, R.M.: Evolutionary games and spatial chaos, Nature 359 (1992) 826-29
11. Nowak, M.A., and Sigmund, K.: A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game, Nature 364 (1993) 56-58
12. Stanley, E.A., Ashlock, D., and Tesfatsion, L.: Iterated prisoner's dilemma with choice and refusal of partners, Artificial Life III, Addison-Wesley (1994)
13. Tzafestas, E.: Toward adaptive cooperative behavior, Proceedings of the Simulation of Adaptive Behavior Conference, Paris, September (2000)
14. Tzafestas, E.: Spatial games with adaptive tit-for-tats, Proceedings of the Simulation on Parallel Problem Solving from Nature (PPSN), Paris, September (2000)
15. Tzafestas, E.: Attraction and cooperation in space, in press.
16. Tzafestas, E.: Emergence of social networks in systems with attraction, submitted.