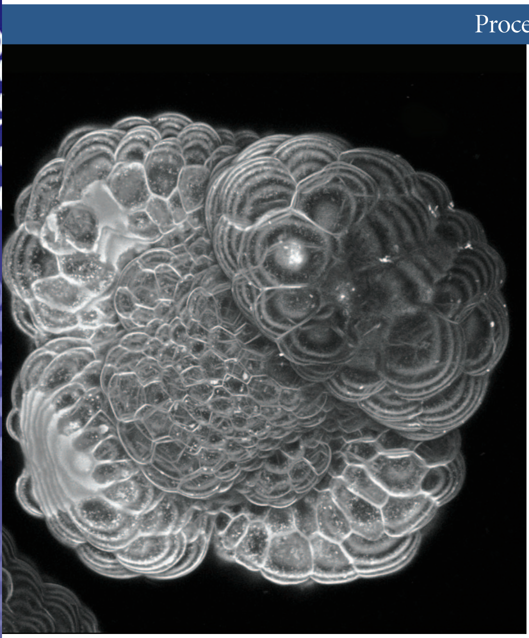
Evolution of Cooperation by Phenotypic Similarity

Citation	Antal Tibor, Hisashi Ohtsuki, John Wakeley, Peter D. Taylor, and Martin A. Nowak. 2009. Evolution of cooperation by phenotypic similarity. Proceedings of the National Academy of Sciences USA 106(21): 8597-8600.
Published Version	doi:10.1073/pnas.0902528106
Accessed	February 18, 2015 10:17:13 AM EST
Citable Link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:4316891
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

May 26, 2009 | vol. 106 | no. 21 | 8401–8792

PNAS

Proceedings of the National Academy of Sciences of the United States of America www.pnas.org



Cover image: When treated with the herbicide oryzalin, the isotropic growth of *Arabidopsis thaliana* results in a froth-like cellular structure. In plants, the resistance of the cell wall to internal turgor pressure helps shape the cells and the tissues, providing plant rigidity and helping it stand erect. Oryzalin depolymerizes microtubules in the growing shoot apical meristem of *A. thaliana*, resulting in isotropic growth as the cell walls yield under turgor pressure. See the article by Francis Corson et al. on pages 8453–8458. Photo by Olivier Hamant.

From the Cover

Contents

THIS WEEK IN PNAS

LETTERS (ONLINE ONLY)

COMMENTARIES

☀ **Free online through the PNAS open access option.**

PROFILE

INAUGURAL ARTICLE

PHYSICAL SCIENCES

APPLIED MATHEMATICS

APPLIED PHYSICAL SCIENCES

Evolution of cooperation by phenotypic similarity

Tibor Antal[a], Hisashi Ohtsuki[b,c], John Wakeley[d], Peter D. Taylor[e], and Martin A. Nowak[a,d,1]

[a]Program for Evolutionary Dynamics and Department of Mathematics, Harvard University, Cambridge, MA 02138; [b]Department of Value and Decision Science, Tokyo Institute of Technology, Tokyo 152-8552, Japan; [c]Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, Saitama 332-0012, Japan; [d]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138; and [e]Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada K7L 3N6

Communicated by Simon A. Levin, Princeton University, Princeton, NJ, March 10, 2009 (received for review June 15, 2008)

The emergence of cooperation in populations of selfish individuals is a fascinating topic that has inspired much work in theoretical biology. Here, we study the evolution of cooperation in a model where individuals are characterized by phenotypic properties that are visible to others. The population is well mixed in the sense that everyone is equally likely to interact with everyone else, but the behavioral strategies can depend on distance in phenotype space. We study the interaction of cooperators and defectors. In our model, cooperators cooperate with those who are similar and defect otherwise. Defectors always defect. Individuals mutate to nearby phenotypes, which generates a random walk of the population in phenotype space. Our analysis brings together ideas from coalescence theory and evolutionary game dynamics. We obtain a precise condition for natural selection to favor cooperators over defectors. Cooperation is favored when the phenotypic mutation rate is large and the strategy mutation rate is small. In the optimal case for cooperators, in a one-dimensional phenotype space and for large population size, the critical benefit-to-cost ratio is given by $b/c = 1 + 2/\sqrt{3}$. We also derive the fundamental condition for any two-strategy symmetric game and consider high-dimensional phenotype spaces.

coalescent theory | evolutionary dynamics | evolutionary game theory | mathematical biology | stochastic process

Evolutionary game theory is the study of frequency-dependent selection (1–8). Fitness values depend on the relative abundance, or frequency, of various strategies in the population, for example, the frequency of cooperators and defectors. Evolutionary game theory has been applied to understand the evolution of cooperative interactions in viruses, bacteria, plants, animals, and humans (9–13). The classical approach to evolutionary game dynamics assumes well-mixed populations, where every individual is equally likely to interact with every other individual (4). Recent advances include the extension to populations that are structured by geography or other factors (14–25).

The term "greenbeard effect" was coined in sociobiology to describe the result of the following thought experiment (26, 27). What evolutionary dynamics will occur if a single gene is responsible for both a phenotypic signal ("a green beard") and a behavioral response (for example, altruistic behavior toward individuals with like phenotypes)? Later, the term "armpit effect" was introduced (28) to refer to a self-referent phenotype that is used in identifying kin (29–31).

Both of these concepts are now seen as cases of "tag-based cooperation," in which a generic system of phenotypic tags is used to indicate similarity or difference, and the evolutionary dynamics of cooperation are studied in the context of these tags. A first approach, based on computer simulations, assumed a well-mixed population, a continuum of tags, and an evolving threshold distance for cooperation (32). More recent models use numerical and analytic methods and often combine tags with viscous population structure (33–37). A general finding of these articles is that it is difficult to obtain cooperation in tag-based models for well-mixed populations, indicating that some spatial structure is needed (14).

Inspired by work on tag-based cooperation (32–34, 38) and building on a previous approach (39), we study evolutionary game dynamics in a model where the behavior depends on phenotypic distance (40, 41). As a particular example we explore the evolution of cooperation (42, 43). Studies of different organisms, including humans, support the idea that cooperation is more likely among similar individuals (31, 44–49). Our model applies to situations where individuals tend to like those who have similar attitudes and beliefs. We introduce a natural model in which individuals mutate to adjacent phenotypes in a possibly multidimensional phenotype space. We study one and infinitely many dimensions in detail. We develop a theory for general evolutionary games, not just the evolution of cooperation. Spatial structure is not needed for cooperation to be favored in our model. Moreover, in contrast to previous work (39), we develop an analytic machinery for describing heterogeneous populations in phenotype space.

Consider a population of asexual haploid individuals, with a population size N that is constant over time. Each individual is characterized by a phenotype, given by an integer i that can take any value from minus to plus infinity. Thus, this phenotype space is a one-dimensional and unbounded lattice. Individuals inherit the phenotype of their parent subject to some small variation. If the parent's phenotype is i, then the offspring has phenotype $i-1$, i, or $i+1$ with probabilities v, $1-2v$, and v, respectively. The parameter v can vary between 0 and 1/2.

Let us consider a Wright–Fisher process. In each generation, all individuals produce the same large number of offspring. The next generation of N individuals is sampled from this pool of offspring. To introduce some fundamental concepts and quantities, we first study the model without any selection. No evolutionary game is yet being played, and there is only neutral drift in phenotype space. The entire population performs a random walk with a diffusion coefficient v, and by this process will tend to disperse over the lattice. In opposition to this, all of the individuals in the population will be, to some degree, related due to reproduction in a finite population. Thus, while occasionally the population may break up into two or more clusters, typically there is only a single cluster (50, 51). The standard deviation of the distribution in phenotype space, which is a measure for the width of the cluster, is $\sqrt{2Nv}$.

Next, we superimpose the neutral drift of two types: the strategies A and B (Fig. 1). Still for the moment, assuming no fitness differences, we have reproduction subject to mutation between A and B. Specifically, with probability u the offspring adopts a random strategy. The mutation–reproduction process defines a stationary distribution (52). If u is very small relative to N, the population tends to be either all-A or all-B. If u is large, the population tends toward one-half A and one-half B. Fig. 2 illustrates the random walk in phenotype space of the population composed of the two types A and B.

Author contributions: T.A., H.O., J.W., P.D.T., and M.A.N. performed research; and T.A., H.O., J.W., P.D.T., and M.A.N. wrote the paper.

The authors declare no conflict of interest.

See Commentary on page 8405.

[1] To whom correspondence should be addressed. E-mail: martin_nowak@harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0902528106/DCSupplemental.

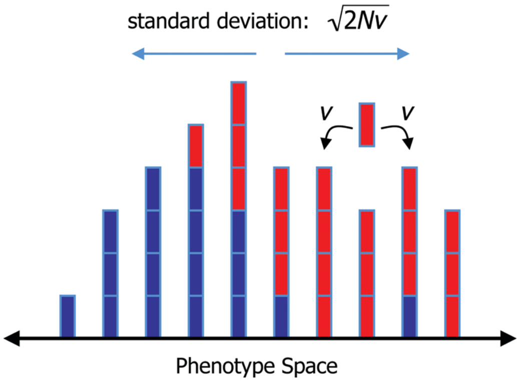


Fig. 1. The basic geometry of evolution in phenotype space. There are two types of individuals (red and blue), which can refer to arbitrary traits or different strategies in an evolutionary game. Individuals inherit the strategy of their parent subject to a small mutation rate u. Moreover, each individual has a phenotype. Here, we consider a discrete one-dimensional phenotype space. An individual of phenotype i produces offspring of phenotype $i-1$, i, or $i+1$ with probabilities v, $1-2v$, and v, respectively. The total population (of size N) performs a random walk in phenotype space with diffusion coefficient v. Sometimes the cluster breaks into two or more pieces, but typically only one of them survives. If evolutionary updating occurs according to a Wright–Fisher process then the distribution of individuals in phenotype space has a standard deviation of $\sqrt{2Nv}$. For the Moran process, the standard deviation is reduced to $\sqrt{Nv}$.

By using coalescence theory (53, 54) many interesting and relevant properties of the distributions of both the strategies and phenotypic tags can be calculated. For example, the probability that two randomly chosen individuals have the same phenotype is $z = 1/\sqrt{8Nv}$. The probability that two randomly chosen individuals have the same strategy and the same phenotype is $g = z(1 - Nu/2)$. The probability that two individuals have the same strategy and a third individual has the same phenotype as the second is $h = z[1 - Nu(2+\sqrt{3})/4]$. These results hold for large population size N and small mutation rate u; more precisely, we assume large Nv and small Nu. The relevance of z, g, and h will become clear below. The expressions for z, g, and h are derived for general Nv and Nu in supporting information (SI) *Appendix*, where they appear as Eqs. **10**, **19**, and **24**, respectively.

We can now use these insights to study game dynamics. We investigate the competition of cooperators, C, and defectors, D. Cooperators play a conditional strategy: they cooperate with all individuals who are close enough in phenotype space and defect otherwise. The notion of being close enough is modeled by a lattice structure. In particular, a cooperator with phenotype i cooperates only with other individuals of phenotype i. Defectors, in contrast, play an unconditional strategy: they always defect. Cooperation means paying a cost, c, for the other individual to receive a benefit b. The larger the total payoff of an individual, interacting equally with every member of the population, the larger the number of offspring it will produce on average. We want to calculate the critical benefit-to-cost ratio, b/c, that allows the game in phenotype space to favor the evolution of cooperation.

A configuration of the population is specified by m_i and n_i, which are the number of cooperators with phenotype i and the total number of individuals with phenotype i, respectively. The total payoff of all cooperators is $F_C = \sum_i m_i(bm_i - cn_i)$. The total payoff of all defectors is $F_D = \sum_i (n_i - m_i)bm_i$. There are $\sum_i m_i$ cooperators and $N - \sum_i m_i$ defectors. The average payoff for a cooperator is $f_C = F_C/\sum_i m_i$. The average payoff for a defector is $f_D = F_D/(N - \sum_i m_i)$. Cooperators have a higher fitness than defectors if $f_C > f_D$, which leads to $\sum_i m_i(bm_i - cn_i) > \sum_i m_i \sum_j m_j n_j (b-c)/N$. Averaging these quantities over every possible configuration of the population, weighted by their stationary probability under neutrality, we obtain the fundamental condition

$$b\left\langle \sum_i m_i^2 \right\rangle - c\left\langle \sum_i m_i n_i \right\rangle > (b-c)\left\langle \sum_{ij} m_i m_j n_j \right\rangle /N. \qquad [1]$$

Under this condition cooperators are more abundant than defectors in the mutation-selection process. The above argument and our results are valid in the weak selection limit. A precise derivation of this inequality is presented in *SI Appendix*. Correlation terms similar to the ones above sometimes arise in studies of social behavior and population dynamics (26, 55). The first two terms in inequality (Eq. **1**) are pairwise correlations, while the third is notably a triplet correlation. Note that the argument leading to inequality (Eq. **1**) includes self-interaction, but that the effect of this becomes negligible when N is large.

When the population size is large, the averages in inequality (Eq. **1**) are proportional to the probabilities g, z, and h respectively, which we introduced earlier. Consequently, inequality (Eq. **1**) can be written as $bg - cz > (b-c)h$. Using the values of z, g, and h given above we obtain

$$b/c > 1 + \frac{2}{\sqrt{3}}, \qquad [2]$$

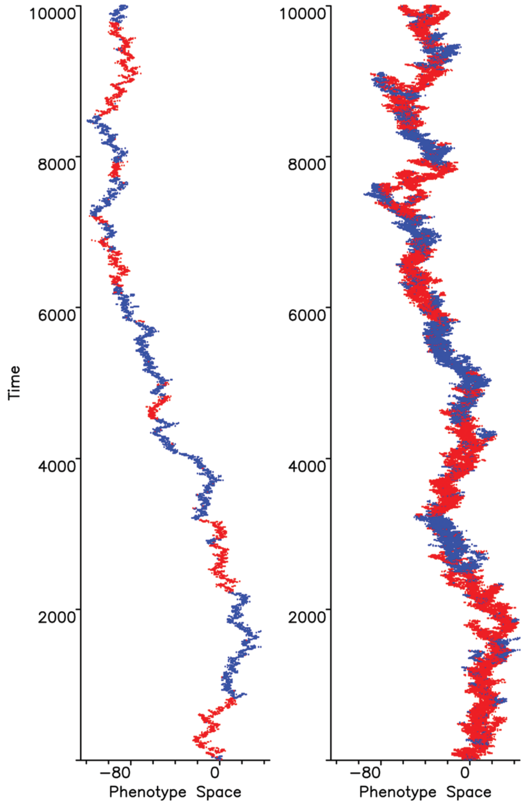


Fig. 2. Random walks in phenotype space. Shown are two computer simulations of a Wright–Fisher process in a one-dimensional discrete phenotype space. The phenotypic mutation rate is $v = 0.25$. The colors, red and blue, refer to arbitrary traits, because no game is yet being played. All individuals have the same fitness. The population size is (*Left*) $N = 10$, and (*Right*) $N = 100$. The strategy mutation probability (between red and blue) is $u = 0.004$. Therefore, a given color dominates on average for $2/u = 500$ generations (since new mutations arrive at rate $Nu/2$ and fixate with probability $1/N$). The standard deviation of the distribution in phenotype space is $\sqrt{2Nv}$. Approximately 95% of all individuals are within 4 standard deviations. Often the population fragments into two or several pieces, but only one branch survives in the long run. We use the statistics of these neutral "phenotypic space walks" for calculating the fundamental conditions of evolutionary games in the limit of weak selection.

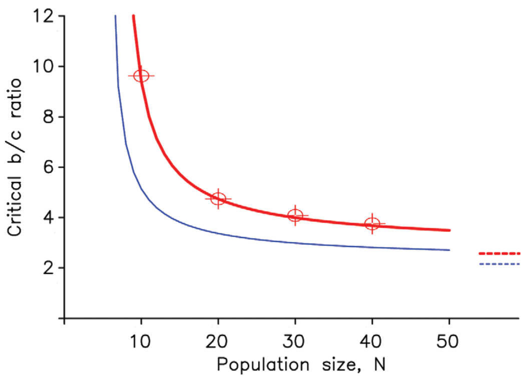


Fig. 3. Excellent agreement between numerical simulations and analytic calculations. We show the critical benefit-to-cost ratio that is needed for cooperators to be more abundant than defectors in the stationary distribution. We have used a Wright–Fisher process with a phenotypic mutation rate $v = 1/2$ and a strategy mutation probability $u = 1/(2N)$. The red line indicates the result of our analytic calculation. For these parameter values the asymptotic limit for large N is $b/c = (1 + 12\sqrt{2})/7 \approx 2.5672$. The red dots indicate the result of numerical simulations. The gray line illustrates the critical b/c-ratio for $u \to 0$ with the asymptotic limit $b/c = 1 + 2/\sqrt{3} \approx 2.1547$.

which is approximately 2.16. If the benefit-to-cost ratio exceeds this number, then cooperators are more abundant than defectors in the mutation-selection process. The success of cooperators results from the balance of movement and clustering in phenotype space. Inequality (Eq. **2**) represents the condition for cooperators to be more abundant than defectors in a large population when the strategy mutation rate u is small ($Nu \ll 1$) and the phenotypic mutation rate v is large ($Nv \gg 1$). In *SI Appendix*, we derive conditions for any population size and mutation rates. Fig. 3 shows the excellent agreement between numerical simulations and analytical calculations. In general, we find that both lowering strategy mutations and increasing phenotypic mutations favor cooperators.

We can expand our analysis to study any 2 × 2 game, not only the interaction between cooperators and defectors. Consider two strategies A and B and the general payoff matrix

$$\begin{array}{cc} & \begin{array}{cc} A & B \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{pmatrix} R & S \\ T & P \end{pmatrix}. \end{array} \quad [3]$$

The payoffs for A versus A, A versus B, B versus A, and B versus B are given by R, S, T, P, respectively. A players use strategy A against other individuals with the same phenotype, otherwise they use B. B players always use strategy B. For the game in a one-dimensional phenotype space and large population size we find that A is more abundant than B if

$$(R - P)(1 + \sqrt{3}) > T - S. \quad [4]$$

For the derivation see Section 5.2 in the *SI Appendix*. This formula can be used for evaluating any two-strategy symmetric game in a one-dimensional phenotype space. In the *SI Appendix*, we discuss the snow-drift game and the stag-hunt game as particular examples.

We can also study higher-dimensional phenotype spaces. In general, for higher dimensions, it is easier for cooperators to overcome defectors. The intuitive reason is that in higher dimensions phenotypic identity also implies strategic identity. In Section 5.3 of the *SI Appendix*, we show that, in the limit of infinitely many dimensions, and under the same assumptions that produced conditions **2** and **4**, the crucial benefit-to-cost ratio in the Prisoner's Dilemma converges to $b/c > 1$. For general games, the equivalent result of condition **4** becomes $R > P$, which means the evolutionary process always chooses the strategy with the higher payoff against itself. Our basic approach can also be adapted to continuous, rather than discrete, phenotype spaces. In this case, no two individuals have exactly the same phenotype, but the conditional behavioral strategy is triggered by sufficient phenotypic similarity.

In summary, we have developed a model for the evolution of cooperation based on phenotypic similarity. Our approach builds on previous ideas of tag-based cooperation, but in contrast to earlier work (33–37), we do not need spatial population dynamics to obtain an advantage for cooperators. We derive a completely analytic theory that provides general insights. We find that the abundance of cooperators in the mutation-selection equilibrium is an increasing function of the phenotypic mutation rate and a decreasing function of the strategic mutation rate. These observations agree with the basic intuition that higher phenotypic mutation rates reduce the interactions between cooperators and defectors, whereas higher strategic mutation rates destabilize clusters of cooperators by allowing frequent invasion of newly mutated defectors. Therefore, cooperation is more likely to evolve if the strategy mutation rate is small and if the phenotypic mutation rate is large. In a genetic model this assumption may be fulfilled if the strategy is encoded by one or a few genes, whereas the phenotype is encoded by many genes. Also in a cultural model, it can be the case that the phenotypic mutation rates are higher than the strategic mutation rates; for example, people might find it easier to modify their superficial appearance than their fundamental behaviors. Furthermore, we show how the correlations between strategies and phenotypes can be obtained from neutral coalescence theory under the assumption that selection is weak (54, 56). Our theory can be applied to study any evolutionary game in the context of conditional behavior that is based on phenotypic similarity or difference.

ACKNOWLEDGMENTS. This work was supported by the John Templeton Foundation, the National Science Foundation/National Institutes of Health (R01GM078986) joint program in mathematical biology, the Bill and Melinda Gates Foundation (Grand Challenges Grant 37874), the Japan Society for the Promotion of Science, and J. Epstein.

1. Smith JM, Price GR (1973) The logic of animal conflict. *Nature* 246:15–18.
2. Taylor PD, Jonker L (1978) Evolutionarily stable strategies and game dynamics. *Math Biosci* 40:145–156.
3. Smith JM (1982) *Evolution and the Theory of Games* (Cambridge Univ Press, Cambridge, UK).
4. Hofbauer J, Sigmund K (1998) *Evoltuionary Games and Population Dynamics* (Cambridge Univ Press, Cambridge, UK).
5. Cressman R (2003) *Evolutionary Dynamics and Extensive Form Games* (MIT Press, Cambridge, MA).
6. Vincent TL, Brown JS (2005) *Evolutionary Game Theory, Natural Selection, and Darwinian Dynamics* (Cambridge Univ Press, Cambridge, UK).
7. Nowak MA, Sigmund K (2004) Evolutionary dynamics of biological games. *Science* 303:793–799.
8. May RM (1973) *Stability and Complexity in Model Ecosystems* (Princeton Univ Press, Princeton, NJ).
9. Parker GA (1974) Assessment strategy and evolution of fighting behavior. *J Theor Biol* 47:223–243.
10. Colman AM (1995) *Game Theory and Its Applications in the Social and Biological Sciences* (Butterworth–Heinemann, Oxford).
11. Sinervo B, Lively CM (1996) The rock-paper-scissors game and the evolution of alternative male strategies. *Nature* 380:240–243.
12. Nee S (2000) Mutualism, parasitism and competition in the evolution of coviruses. *Philos Trans R Soc London Ser B* 355:1607–1613.
13. Kerr B, Riley MA, Feldman MW, Bohannan BJ (2002) Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature* 418:171–174.
14. Nowak MA, May RM (1992) Evolutionary games and spatial chaos. *Nature* 359:826–829.
15. Durrett R, Levin SA (1994) The importance of being discrete (and spatial). *Theor Popul Biol* 46:363–394.
16. Hassell MP, Comins HN, May RM (1994) Species coexistence and self-organizing spatial dynamics. *Nature* 370:290–292.
17. Killingback T, Doebeli M (1996) Spatial evolutionary game theory: Hawks and Doves revisited. *Proc R Soc London Ser B* 263:1135–1144.
18. Nakamaru M, Matsuda H, Iwasa Y (1997) The evolution of cooperation in a lattice-structured population. *J Theor Biol* 184:65–81.
19. Eshel I, Sansone E, Shaked A (1999) The emergence of kinship behavior in structured populations of unrelated individuals. *Int J Game Theory* 28:447.
20. Neuhauser C, Pacala S (1999) An explicitly spatial version of the Lotka-Volterra model with interspecific competition. *Ann Appl Prob* 9:1226–1259.

21. Szabo G, Hauert C (2002) Phase transitions and volunteering in spatial public goods games. *Phys Rev Lett* 89:118101.
22. Hauert C, Doebeli M (2004) Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature* 428:643–646.
23. Ohtsuki H, Hauert C, Lieberman E, Nowak MA (2006) A simple rule for the evolution of cooperation on graphs and social networks. *Nature* 441:502–505.
24. Santos FC, Pacheco JM, Lenaerts T (2006) Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proc Natl Acad Sci USA* 103:3490–3494.
25. Taylor PD, Day T, Wild G (2007) Evolution of cooperation in a finite homogeneous graph. *Nature* 447:469–472.
26. Hamilton WD (1964) The genetical behavior of social behavior I. *J Theor Biol* 7:1–16.
27. Dawkins R (1976) *The Selfish Gene* (Oxford Univ Press, Oxford).
28. Dawkins R (1982) *The Extended Phenotype* (Oxford Univ Press, Oxford).
29. Matteo JM, Johnston RE (2000) *Proc R Soc London* 267:695–700.
30. Sinervo B, et al. (2006) Self-recognition, color signals, and cycles of greenbeard mutualism and altruism. *Proc Natl Acad Sci USA* 103:7372–7377.
31. Lize A, et al. (2006) Kin discrimination and altruism in the larvae of a solitary insect. *Proc R Soc London Ser B* 273:2381–2386.
32. Riolo RL, Cohen MD, Axelrod R (2001) Evolution of cooperation without reciprocity. *Nature* 414:441–443.
33. Axelrod R, Hammond RA, Grafen A (2004) Altruism via kin-selection strategies that rely on arbitrary tags with which they coevolve. *Evolution (Lawrence, Kans)* 58:1833–1838.
34. Jansen VAA, van Baalen M (2006) Altruism through beard chromodynamics. *Nature* 440:663–666.
35. Hammond RA, Axelrod R (2006) Evolution of contingent altruism when cooperation is expensive. *Theor Pop Biol* 69:333–338.
36. Rousset F, Roze D (2007) Constraints on the origin and maintenance of genetic kin recognition. *Evolution (Lawrence, Kans)* 61:2320–2330.
37. Gardner A, West SA (2007) Social evolution: The decline and fall of genetic kin recognition. *Curr Biol* 17:R810–R812.
38. Hochberg ME, Sinervo B, Brown SP (2003) Socially mediated speciation. *Evolution (Lawrence, Kans)* 57:154–158.
39. Traulsen A, Nowak MA (2007) Chromodynamics of cooperation in finite populations. *PLoS ONE* 2:e270.
40. Levin SA, Segel LA (1982) Models of the influence of predation on aspect diversity in prey populations. *J Math Biol* 14:253–284.
41. Levin SA, Segel LA (1985) Pattern generation in space and aspect. *SIAM Rev* 27:45–67.
42. Axelrod R, Hamilton WD (1981) The evolution of cooperation. *Science* 211:1390–1396.
43. Nowak MA (2006) Five rules for the evolution of cooperation. *Science* 314:1560–1563.
44. Byrne D (1969) Attitudes and attraction. *Adv Exp Soc Psychol* 4:35–89.
45. Nahemow L, Lawton MP (1975) Similarity and propinquity in friendship formation. *J Personal Soc Psychol* 32:205–213.
46. Selfhout MHW, et al. (2007) The role of music preferences in early adolescents' friendship formation and stability. *J Adolesc* 32:95–107.
47. Tajfel H, Billig RP, Flament C (1971) Social categorization and intergroup behavior. *Eur J Soc Psychol* 1:149–178.
48. Burger JM, Messian N, Patel S, Prado AD, Anderson C (2004) What a coincidence! The effects of incidental similarity on compliance. *Personality Soc Psychol Bull* 30:35–43.
49. Rand DG, et al. (2009) Dynamic remodeling of in-group bias during the 2008 presidential election. *Proc Natl Acad Sci USA*, 10.1073/pnas.0811552106.
50. Moran PAP (1975) Wandering distributions and electrophoretic profile. *Theor Popul Biol* 8:318–330.
51. Kingman JFC (1976) Coherent random-walks arising in some genetic models. *Proc R Soc London Ser A* 351:19–31.
52. Wright S (1931) Evolution in Mendelian populations. *Genetics* 16:97–159.
53. Kingman JFC (1982) On the genealogy of large populations. *J Appl Prob* 19:27–43.
54. Wakeley J (2008) *Coalescent Theory: An Introduction* (Roberts & Company Publishers, Greenwood Village, CO).
55. Price GR (1970) Selection and covariance. *Nature* 227:520–521.
56. Rousset F (2003) A minimal derivation of convergence stability measures. *J Theor Biol* 221:665–668.

APPENDIX FOR “EVOLUTION OF COOPERATION BY PHENOTYPIC SIMILARITY”

TIBOR ANTAL, HISASHI OHTSUKI, JOHN WAKELEY, PETER D. TAYLOR, MARTIN A. NOWAK

Contents

1. Model

Consider a population of N haploid individuals (players). Each individual $k = 1, \dots, N$ has an integer-valued phenotype $X_k \in \mathbb{Z}$, which we also refer to as its position in phenotype space. Additionally, each individual has a strategy $S_k \in \{0, 1\}$, and we refer to these two strategies as cooperation (1) and defection (0). In general, players' phenotypes and strategies determine their fitness.

We study the Wright-Fisher (W-F) process, where each of the N individuals of the next generation independently chooses a parent from the previous generation with a probability proportional to the parent's fitness. Each offspring inherits the parent's position (phenotype) with probability $1 - 2v$, and it is placed to either the left or the right neighboring position of the parent, both with probability v. Each offspring also inherits the parent's strategy with probability $1 - u$, and it adopts a random strategy with probability u.

We derive the condition for cooperation to be favored in the large population size limit. This condition depends on certain correlations in the neutral case, that is when each individual has the same fitness. These correlations are calculated in Section 2. Then in Section 3 the condition for cooperation is derived. In Section 4 we discuss finite population sizes, cooperation without self interaction, and a precise derivation of the correlations. Finally in Section 5 as generalizations of our model we consider the Moran process, general payoff matrices, and we discuss an infinite dimensional phenotype space.

2. Correlations in the neutral case

In this section we consider the neutral case, that is when all players have the same fitness. Note that the strategies and the phenotypes of the individuals change independently, and evolve according to the Wright-Fisher process [1, 2]. The system rapidly reaches a stationary state where the individuals stay in a cluster with variance $2Nv$, but the cluster as a whole diffuses over the space (the integers) with diffusion coefficient v. We are interested in the properties of this stationary state.

We are particularly interested in four probabilities. We pick three distinct individuals k, q, and l from the population in the stationary state. For their phenotypes and their strategies we define the following four probabilities

$$\begin{aligned} y &= \Pr(S_k = S_q) \\ z &= \Pr(X_k = X_q) \\ g &= \Pr(S_k = S_q,\ X_k = X_q) \\ h &= \Pr(S_l = S_k,\ X_k = X_q) \end{aligned} \tag{1}$$

In words, y is the probability that two individuals have the same strategy, and z is the probability that they have the same phenotype. They have simultaneously the same strategy and phenotype with probability g. Out of three individuals, the probability that the first two have the same phenotype, and simultaneously the first and the third have the same strategy is denoted by h. Note that neither g nor h factorizes in general.

To obtain the above probabilities we have to know the probability $\Pr(T = t)$ that the time T to the most recent common ancestor (MRCA) of two randomly chosen individual is $T = t$. This time is not affected by either the strategies or the phenotypes of the players. It is determined solely by the W-F dynamics. The ancestry of two individuals coalesce with probability $1/N$ in each time step. Hence the probability that the time to the MRCA is t is

$$\Pr(T = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N} \tag{2}$$

We can continue the calculation for finite system size N, but the expressions become cumbersome. Hence we relegated the finite N calculations to Section 4.1, where we mainly treat the special $v = 1/2$ case. In this section we discuss the large population limit $N \to \infty$, where we introduce the rescaled time $\tau = t/N$. In this limit we can use a continuous time description, where the coalescent time distribution (2) is given by the density function

$$p(\tau) = e^{-\tau} \tag{3}$$

and the average coalescence time becomes $\tau = 1$ in the new unit.

Due to the non-overlapping generations in the W-F model, each individual is a newborn and has the chance to mutate both in strategy and phenotype space. In the large N and

$u, v \to 0$ limit, the system can be described as a continuous time process. Strategy mutations arrive at rate $\mu = 2Nu$ and phenotype mutations at rate $\nu = 2Nv$ (in each direction) on the ancestral line of *two* individuals. Note that this continuous time limit is exact for the Moran process even for finite values of v, as it is shown in Section 5.1. In the W-F model, for finite values of v we have a discrete time random walk, but the typical number of steps goes to infinity. In that limit the discrete and continuous time walks become identical, and hence the finite v behavior can be recovered as the $\nu \to \infty$ limit.

2.1. **Phenotypic distance.** Let us first study the phenotypes of the players. Here we calculate not only z, but in general the probability that two randomly chosen individuals k and q are at distance x in phenotype space

$$z(x) = \Pr(X_k - X_q = x) \tag{4}$$

We know that the (signed) distance between the two individuals changes by plus or minus one at rate ν, and the distance distribution after time τ can be expressed in terms of the Modified Bessel functions [3, 4] as

$$\zeta(x|\tau) = e^{-2\nu\tau} I_{|x|}(2\nu\tau) \tag{5}$$

The probability that two individuals are distance x apart is

$$z(x) = \sum_{t=1}^{\infty} \Pr(X_k - X_q = x | T = t) \Pr(T = t) \tag{6}$$

which becomes an integral of the corresponding density functions in the continuous time limit

$$z(x) = \int_0^\infty p(\tau)\zeta(x|\tau) d\tau = \int_0^\infty e^{-(2\nu+1)\tau} I_{|x|}(2\nu\tau) d\tau \tag{7}$$

By using the identity [5]

$$\int_0^\infty e^{-ac} I_\gamma(bc)\, dc = \frac{b^{-\gamma}\left(a - \sqrt{a^2 - b^2}\right)^\gamma}{\sqrt{a^2 - b^2}} \tag{8}$$

we arrive at the probability distribution of the signed distance

$$z(x) = \frac{1}{\sqrt{4\nu + 1}} \left(\frac{2\nu + 1 - \sqrt{4\nu + 1}}{2\nu} \right)^{|x|} \tag{9}$$

The individuals are at the same position with probability

$$z \equiv z(0) = \frac{1}{\sqrt{4\nu + 1}} \tag{10}$$

Distribution (9) is of course normalized $\sum_{x=-\infty}^{\infty} z(x) = 1$, and its second moment is

$$\sum_{x=-\infty}^{\infty} x^2 z(x) = 2\nu \tag{11}$$

Note that this second moment is twice the variance of the individual positions, which is exactly $\nu = 2Nv$ even for finite N (see Section 4.1). Hence the individuals stay together in a cluster of size $\sqrt{2Nv}$. This cluster diffuses collectively through phenotype space. If one follows the ancestral line of an individual time τ back, its position $\hat{x}(\tau)$ will change by one at rate $\nu/2$ in each direction. Consequently, the position of the cluster has a variance proportional to time

$$\langle \hat{x}^2 \rangle = \nu\tau = 2vt \tag{12}$$

which implies a diffusive motion. The same result is valid for any finite N in the large time limit. Note that the diffusion coefficient $D = v$ does not depend on the population size. Since the cluster itself wanders in space, the average number of individuals at any given site goes to zero. That is why we focus on distances in the phenotype space (4).

2.2. **Pair with same strategy.** We are interested in the probability y that two randomly chosen individuals have the same strategy. In the continuous time limit, strategy mutations arrive at rate μ on the ancestral lines of the two individuals. The two individuals have the same strategy if there were no mutations, which is the case with probability $e^{-\mu\tau}$. Otherwise there was at least one mutation, hence at least one of the players has a random strategy, so they have the same strategy with probability 1/2. Consequently, the probability that two players have the same strategy time τ after their MRCA is

$$y(\tau) = e^{-\mu\tau} + \frac{1}{2}\left(1 - e^{-\mu\tau}\right) \tag{13}$$

The probability y that two randomly chosen individuals have the same strategy is

$$y = \sum_{t=1}^{\infty} \Pr(S_k = S_q | T = t) \Pr(T = t) \tag{14}$$

In the continuous time limit we obtain

$$y = \int_0^{\infty} p(\tau) y(\tau) d\tau = \frac{2+\mu}{2(1+\mu)} \tag{15}$$

where we have used (3) and (13).

2.3. **Pair with same strategy and phenotype.** The probability g that two randomly chosen individuals have the same phenotype and also have the same strategy can be obtained as

$$g = \sum_{t=1}^{\infty} \Pr(S_k = S_q | T = t) \Pr(X_k = X_q | T = t) \Pr(T = t) \tag{16}$$

Here we have used the property, that although g does not factorize in general, nevertheless for any given time t the conditional probabilities factorize as

$$\Pr(S_k = S_q,\ X_k = X_q | T = t) = \Pr(S_k = S_q | T = t) \Pr(X_k = X_q | T = t) \tag{17}$$

The reason is that mutations occur completely independently in the strategy and the phenotype space. The corresponding integral in the continuous time limit hence becomes

$$g = \int_0^{\infty} p(\tau) y(\tau) \zeta(\tau) d\tau \tag{18}$$

where we use the notation $\zeta(\tau) \equiv \zeta(0|\tau)$. Note that it is also easy to obtain the analog probability where the phenotype difference is x, but we do not consider that here. Using identity (8) again, we can evaluate the above integral

$$g = \frac{1}{2\sqrt{1+4\nu}} + \frac{1}{2\sqrt{(1+\mu)(1+\mu+4\nu)}} \tag{19}$$

2.4. **Three point correlations.** Now we turn to the calculation of the three point probability h which is defined in (1). If we follow the ancestral lines of three individuals back in time, the probability that there was no coalescence event during one update step is $(1-1/N)(1-2/N)$. Two individuals coalesce with probability $3/N \cdot (1-1/N)$. When two individual have coalesced, the remaining two coalesce with probability $1/N$ during each update step. Hence the probability that the first merging happens to any pair of individuals at time $t_3 \geq 1$ back in time, and the second $t_2 \geq 1$ before the first one is

$$\Pr(t_3, t_2) = \frac{3}{N^2}\left[\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\right]^{t_3-1}\left(1-\frac{1}{N}\right)^{t_2} \tag{20}$$

The probability that three individual coalesce simultaneously at time t_3 is

$$\Pr(t_3, 0) = \frac{1}{N^2}\left[\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\right]^{t_3-1} \tag{21}$$

In the $N \to \infty$ limit (20) converges to the density function

$$p(\tau_3, \tau_2) = 3e^{-(3\tau_3+\tau_2)} \tag{22}$$

with $\tau_3 = t_3/N$ and $\tau_2 = t_2/N$. Note that (21) does not affect the large N limit.

Let us call the scaled time when individuals q, k coalesce τ_{qk}, and when k, l coalesce τ_{kl}. With probability 1/3 individuals q, k coalesce first at $\tau_{qk} = \tau_3$ and they coalesce with l at $\tau_{kl} = \tau_3 + \tau_2$. Similarly with probability 1/3 individuals k, l coalesce first at $\tau_{kl} = \tau_3$ and they coalesce with q at $\tau_{qk} = \tau_3 + \tau_2$. If, however, l, q coalesce first with probability 1/3, it makes $\tau_{qk} = \tau_{kl} = \tau_3 + \tau_2$. Since we know the probability density $y(\tau)$ that two individuals with a MRCA at time τ back have the same strategy (13), and the probability density $\zeta(\tau) \equiv \zeta(0|\tau)$ that they are at the same position (5), we can simply obtain the three point correlation as

$$h = \frac{1}{3}\int_0^\infty d\tau_3 \int_0^\infty d\tau_2 \; p(\tau_3, \tau_2)\left[\zeta(\tau_3)y(\tau_3+\tau_2) + \zeta(\tau_3+\tau_2)y(\tau_3) + \zeta(\tau_3+\tau_2)y(\tau_3+\tau_2)\right] \tag{23}$$

This integral can be evaluated by first introducing a variable for $\tau_2 + \tau_3$ in the last two terms of the integral, and by using identity (8) in all three terms. We obtain

$$h = \frac{(1+\mu)(3+\mu) + C_1(2+\mu) - \mu C_3}{2(1+\mu)(2+\mu)\sqrt{1+4\nu}} \tag{24}$$

with the shorthand notation

$$C_i = \frac{1}{2}\sqrt{\frac{(i+\mu)(1+4\nu)}{i+\mu+4\nu}} \tag{25}$$

By now we have obtained all the correlations in (1) in the $N \to \infty$ limit for any values of ν and μ.

3. Threshold b/c ratio

In this section the individuals play a simplified Prisoner's Dilemma game given by the payoff matrix

(26)

		when playing against	
		C	D
payoff of	C	$b-c$	$-c$
	D	b	0

Here $b > 0$ is the benefit gained from cooperators, and $c > 0$ is the cost payed by cooperators. We assume that all individuals interact (in this sense the population is "well mixed"). Cooperators, however, play a conditional strategy: they cooperate with other individuals who have the same phenotype, and they defect otherwise. Defectors always defect. The total payoff of an individual is the sum of all payoffs that individual receives. We introduce the effective payoff of an individual $f = 1 + \delta \cdot \text{payoff}$, where $\delta > 0$ is the strength of the selection, and $\delta = 0$ corresponds to the neutral case discussed in Section 2. Note that δ must be sufficiently small to make all fitness values positive.

We consider here the simplest possible case, where each individual also receives a payoff from self interaction. Excluding self-interaction results in a $1/N$ correction, which is discussed in Section 4.2. An extension to a general payoff matrix is considered in Section 5.2.

3.1. Fitness.

Let n_i denote the number of players of phenotype i, and m_i the number of *cooperators* of phenotype i. A state of the system is given by the vectors $s = (\boldsymbol{n}, \boldsymbol{m})$. Let $f_{C,i}$ and $f_{D,i}$ represent the (effective) payoffs of a cooperator and a defector, respectively, of phenotype i. When self interaction is included these values are

$$\begin{aligned} f_{C,i} &= 1 + \delta\,[bm_i - cn_i] \\ f_{D,i} &= 1 + \delta\,[bm_i]\,. \end{aligned} \tag{27}$$

Let $w_{C,i}$ and $w_{D,i}$ represent the fitness (*i.e.* average number of offsprings) of a cooperator and a defector of phenotype i. After one update step (which is one generation) we obtain

$$w_{C,i} = \frac{N f_{C,i}}{\sum_j [m_j f_{C,j} + (n_j - m_j) f_{D,j}]} \tag{28}$$

Here a cooperator is chosen to be a parent with probability given by its payoff relative to the total payoff, and this happens N times independently in one update step. The denominator of (28) can be written as

$$\sum_j [m_j f_{C,j} + (n_j - m_j) f_{D,j}] = N + \delta(b-c) \sum_j m_j n_j \tag{29}$$

Therefore, in the $\delta \to 0$ limit, we obtain the fitness of a phenotype i cooperator

$$w_{C,i} = 1 + \delta \left(bm_i - cn_i - \frac{b-c}{N} \sum_j m_j n_j \right) + \mathcal{O}(\delta^2) \tag{30}$$

3.2. **Effect of selection.** Let p denote the frequency of cooperators in the population. Cooperation is favored if cooperators are in the majority at the stationary state, $\langle p \rangle > 1/2$. The frequency of cooperators p changes during one update step due to selection and due to mutation. In any state s of the system, the total change of cooperator frequency can be expressed in terms of the change due to selection as

$$\Delta p_{\mathrm{tot}}(s) = (1-u)\Delta p_{\mathrm{sel}}(s) + u\left(\frac{1}{2} - p\right) \tag{31}$$

Here the first term describes the change due to selection in the absence of mutation, which happens with probability $1-u$. The second term stands for the effect of mutation, which happens with probability u to each player independently. In this latter case the frequency p increases in average by $1/2$ due to the introduction of random strategies, and decreases by p due to the replacement of cooperators.

In the stationary state $\langle p \rangle$ is constant, hence the total change of frequency vanishes $\langle \Delta p \rangle_{\mathrm{tot}} = 0$. Then from (31) we can express the average cooperator frequency with the change of frequency due to selection as

$$\langle p \rangle = \frac{1}{2} + \frac{1-u}{u}\langle \Delta p \rangle_{\mathrm{sel}} \tag{32}$$

This means that by calculating the average change of cooperator frequency, we also obtain the average cooperator frequency. It also means that cooperators are favored $\langle p \rangle > 1/2$ if their change due to selection is positive in the stationary state

$$\langle \Delta p \rangle_{\mathrm{sel}} > 0 \tag{33}$$

Now let us perform a perturbative expansion for small selection $\delta \ll 1$. In a given state $s = (\boldsymbol{n}, \boldsymbol{m})$, the expected change of p due to selection in one update step is

$$\Delta p(s) = \frac{1}{N}\left(\sum_i m_i w_{C,i} - \sum_i m_i\right) \tag{34}$$

This expression vanishes for $\delta = 0$ for the fitness function (30). (Note that this statement is not true in general for arbitrary models). Its Taylor expansion is

$$\Delta p(s) = 0 + \delta \frac{d\Delta p(s)}{d\delta}\Big|_{\delta=0} + \mathcal{O}(\delta^2) = \frac{\delta}{N}\sum_i m_i \frac{dw_{C,i}}{d\delta}\Big|_{\delta=0} + \mathcal{O}(\delta^2) \tag{35}$$

We also expand the stationary probabilities of finding the system in state s

$$\pi(s) = \pi^{(0)}(s) + \delta\pi^{(1)}(s) + \mathcal{O}(\delta^2) \tag{36}$$

where $\pi^{(0)}(s)$ is the stationary probability in the neutral state (here we consider two states equivalent if they only differ by translation along the phenotype space). Consequently, in the stationary state in the presence of the game, the average change in cooperator frequency can be expressed in the leading order in terms of averages in the neutral stationary state

$$\langle \Delta p \rangle_{\mathrm{sel}} = \frac{\delta}{N}\left\langle \sum_i m_i \frac{dw_{C,i}}{d\delta} \right\rangle_0 + \mathcal{O}(\delta^2) \tag{37}$$

This expression has to be positive for cooperation to be favored (33). Here the 0 subscript refers to $\delta = 0$, that is to an average taken in the stationary state of the neutral model

$\langle\cdot\rangle_0 = \sum_s \cdot\, \pi^{(0)}(s)$. More generally, one can also easily obtain higher order terms in δ based on (35) and (36). The first derivative of the effect of selection in the stationary state

$$\langle \Delta p\rangle^{(1)}_{\text{sel}} = \frac{d\langle \Delta p\rangle_{\text{sel}}}{d\delta}\bigg|_{\delta=0} \tag{38}$$

can be obtained from (37), by using the fitness (30) of our model, as

$$\langle \Delta p\rangle^{(1)}_{\text{sel}} = \frac{1}{N}\left[b\left\langle \sum_i m_i^2 \right\rangle_0 - c\left\langle \sum_i m_i n_i \right\rangle_0 - \frac{b-c}{N}\left\langle \sum_{i,j} m_i m_j n_j \right\rangle_0 \right] \tag{39}$$

The threshold model parameters are then obtained when the change $\langle \Delta p\rangle^{(1)}_{\text{sel}} = 0$, as follows from the general condition (33)

$$\left(\frac{b}{c}\right)^* = \frac{\langle \sum_i m_i n_i\rangle_0 - \frac{1}{N}\left\langle \sum_{i,j} m_i m_j n_j\right\rangle_0}{\langle \sum_i m_i^2\rangle_0 - \frac{1}{N}\left\langle \sum_{i,j} m_i m_j n_j\right\rangle_0} \tag{40}$$

Hence, we have expressed the threshold b/c ratio in the small selection limit in terms of correlations in the neutral stationary state. Note that the averages in (39) cannot be moved inside the sum, since at any given position any stationary average is zero. Also note that all terms in (40) are of order N^2.

The above derivation is valid for finite N and $\delta \to 0$. We are also interested, however, in the $N \to \infty$ asymptotic behavior. In that case all the above derivation can be repeated when simultaneously $\delta N \to 0$.

Expression (39) for the change in cooperator frequency can be rewritten in a more intuitive way. First we express the total payoffs of cooperators and defectors respectively as

$$\begin{aligned} f_C &= \sum_i m_i f_{C,i} = N_C + \delta F_C \\ f_D &= \sum_i m_i f_{D,i} = N_D + \delta F_D \end{aligned} \tag{41}$$

in a given state, where F_C and F_D are the total payoffs without considering weak selection

$$F_C = \sum_i m_i(b m_i - c n_i)\,, \quad F_D = \sum_i (n_i - m_i) b m_i \tag{42}$$

and $N_C = \sum_i m_i$ and $N_D = N - N_C$ are the number of cooperators and defectors respectively. With this notation the change in cooperator frequency (35) can be rewritten as

$$\Delta p(s) = \frac{\delta}{N^2}\left(N_D F_C - N_C F_D\right) + \mathcal{O}(\delta^2) \tag{43}$$

This expression was obtained in an intuitive way in the main text. By averaging over the stationary state we of course recover (39).

3.3. **Threshold value from correlations.** Let us now evaluate the expected values in (40). We randomly choose three individuals k, q, and l with replacement. All expected values in

(40) can be expressed in terms of probabilities in the neutral stationary state

$$\left\langle \sum_i m_i^2 \right\rangle_0 = N^2 \,\Pr(S_k = S_q = 1,\ X_k = X_q) \tag{44a}$$

$$\left\langle \sum_i m_i n_i \right\rangle_0 = N^2 \,\Pr(S_k = 1,\ X_k = X_q) \tag{44b}$$

$$\left\langle \sum_{i,j} m_i m_j n_j \right\rangle_0 = N^3 \,\Pr(S_l = S_k = 1,\ X_k = X_q) \tag{44c}$$

The indices i and j refer to positions, while k, q and l refer to individuals. These identities are self explanatory, nevertheless they are proven in Section 4.3.

Because the two strategies are equivalent in the *neutral* stationary state, all expressions (44) remain valid when we change any 1 to 0. Consequently all expressions (44) simplify to

$$\begin{aligned} \left\langle \sum_i m_i^2 \right\rangle_0 &= \frac{N^2}{2} \,\Pr(S_k = S_q,\ X_k = X_q) \\ \left\langle \sum_i m_i n_i \right\rangle_0 &= \frac{N^2}{2} \,\Pr(X_k = X_q) \\ \left\langle \sum_{i,j} m_i m_j n_j \right\rangle_0 &= \frac{N^3}{2} \,\Pr(S_l = S_k,\ X_k = X_q) \end{aligned} \tag{45}$$

Note that these probabilities are denoted in the main text as P_2, P_1, and P_3 respectively. Substituting the probabilities of (45) into (40) we arrive at the general condition expressed in terms of two and three point correlations

$$\left(\frac{b}{c}\right)^* = \frac{\Pr(S_l = S_k,\ X_k = X_q) - \Pr(X_k = X_q)}{\Pr(S_l = S_k,\ X_k = X_q) - \Pr(S_k = S_q,\ X_k = X_q)} \tag{46}$$

In Section 2 we have calculated similar probabilities defined in (1), but always for two different individuals. In other words while in the probabilities of (45) we pick two individuals with replacement, in the quantities of (1) two individuals were picked without replacement. We know, however, that out of two individuals we pick the same individual twice with probability $1/N$, and pick two different individuals otherwise. We also know the corresponding probabilities when picking three individuals. With this knowledge we can express the probabilities with replacement in (45) with the probabilities without replacement in (1) as follows

$$\begin{aligned} \Pr(S_k = S_q,\ X_k = X_q) &= \frac{1}{N} \,[(N-1)g + 1] \\ \Pr(X_k = X_q) &= \frac{1}{N} \,[(N-1)z + 1] \\ \Pr(S_l = S_k,\ X_k = X_q) &= \frac{1}{N^2} \,[(N-1)(N-2)h + (N-1)\,(z + y + g) + 1] \end{aligned} \tag{47}$$

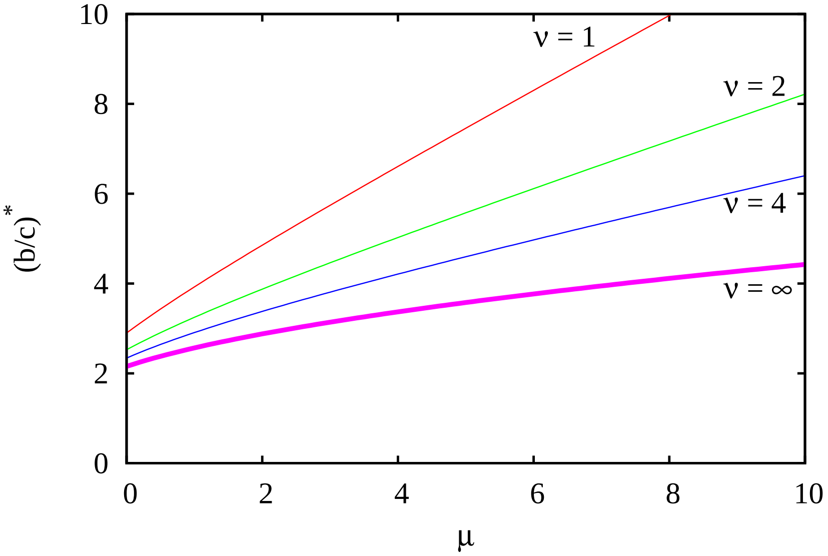


FIGURE 1. Exact threshold b/c ratio (50) in the $N \to \infty$ limit for several values of ν. Cooperation is most favored in the $\mu \to 0$ and $\nu \to \infty$ limit, where $(b/c)^* = 1 + 2/\sqrt{3}$.

Now we substitute these probabilities into condition (46) to obtain the threshold condition

$$\left(\frac{b}{c}\right)^* = \frac{(N-2)(z-h)+1-y+z-g}{(N-2)(g-h)+1-y-z+g} \tag{48}$$

The above condition (48) is exact for any finite N with self interaction. Without self interaction a $\mathcal{O}(1/N)$ correction appears as discussed in Section 4.2. The model of course makes no sense for $N = 1$, and the smallest interesting population size is $N = 2$. In the $N \to \infty$ limit of (48) we also obtain a simple rule

$$\left(\frac{b}{c}\right)^* = \frac{z-h}{g-h} \tag{49}$$

Substituting the expressions (10), (19), and (24) into the above equation for z, g, and h respectively, we arrive at

$$\left(\frac{b}{c}\right)^* = \frac{\mu C_3 - (2+\mu)C_1 + (1+\mu)^2}{\mu C_3 + (2+\mu)C_1 - (1+\mu)} \tag{50}$$

where we have used the shorthand notation (25). This is our main result: the exact threshold b/c ratio in the $N \to \infty$ and weak selection limit. For parameter values $b/c > (b/c)^*$ there are more cooperators than defectors in the system in the long time average.

In Figure 1, we plot the exact $(b/c)^*$ ratio (50) as a function of μ for several values of ν. One observes that $(b/c)^*$ gets smaller both for smaller μ and for larger ν . Hence small strategy mutation and large phenotype mutation helps cooperation. The large ν limit includes the finite v (phenotype changing probability) case. Note that since the cluster size in phenotype space is $\sqrt{2Nv}$, the average number of individuals with the same phenotype is proportional to $\sqrt{N/v}$, hence there are plenty of individuals to interact with even for finite v values in the large N limit.

In the $\nu \to \infty$ limit (50) becomes

$$\left(\frac{b}{c}\right)^* = \frac{(2+\mu)\sqrt{1+\mu} - 2(1+\mu)^2 - \mu\sqrt{3+\mu}}{-(2+\mu)\sqrt{1+\mu} + 2(1+\mu) - \mu\sqrt{3+\mu}} + \mathcal{O}(\frac{1}{\sqrt{\nu}}) \tag{51}$$

which for $\mu \to 0$ behaves as

$$\left(\frac{b}{c}\right)^* = 1 + \frac{2\sqrt{3}}{3} + \mu\frac{7\sqrt{3}-3}{18} + \mathcal{O}(\mu^2) \tag{52}$$

which is ≈ 2.16 in the leading order. For $\mu \to \infty$ the threshold ratio (51) diverges as

$$\left(\frac{b}{c}\right)^* = \sqrt{\mu} + 1 + \mathcal{O}(\frac{1}{\sqrt{\mu}}) \tag{53}$$

Conversely, in the $\mu \to 0$ limit (50) becomes

$$\left(\frac{b}{c}\right)^* = \frac{\sqrt{3}(1+4\nu)^{3/2} + (3+8\nu)\sqrt{3+4\nu}}{\sqrt{3}(1+4\nu)^{3/2} - \sqrt{3+4\nu}} + \mathcal{O}(\mu) \tag{54}$$

This limit function diverges as $3/4\nu$ for small ν, but converges to the constant $1+2/\sqrt{3}$ as $\nu \to \infty$. Hence the best scenario for cooperation is $\mu \to 0$ and $\nu \to \infty$ where $(b/c)^* = 1+2/\sqrt{3}$.

The large N asymptotic results are identical for the Moran process, where we choose a random individual to die, and another (with replacement) to reproduce with probability proportional to the player's payoff (see Section 5.1).

We would like to briefly comment on the relationship between our work and inclusive fitness or kin selection theory (see references in the main text). Let R be the inverse of the r.h.s. of (46). Now we formally obtained Hamilton's rule $(b/c)^* = 1/R$. By dividing both the numerator and the denominator in R by $\Pr(X_k = X_q)$, (we can assume that it is not zero), and using the definition of conditional probability, we can rewrite R as

$$R = \frac{\Pr(S_k = S_q|\ X_k = X_q) - \Pr(S_l = S_k|\ X_k = X_q)}{1 - \Pr(S_l = S_k|\ X_k = X_q)} \tag{55}$$

Now with the notation

$$G = \Pr(S_k = S_q|\ X_k = X_q), \quad \overline{G} = \Pr(S_l = S_k|\ X_k = X_q) \tag{56}$$

we obtain $R = (G - \overline{G})/(1 - \overline{G})$, which is in the form of usual relatedness formula. Note, however, that this $\overline{G}$ is not the probability of identity in state (IIS) between two random individuals in the population as it usually is in inclusive fitness theory. Instead, $\overline{G}$ is a sort of weighted average of IIS probabilities in which those who share the same phenotype with more players are assigned a larger weight.

4. Further clarifications

4.1. **Finite populations for $v = 1/2$.** Here we consider the Wright-Fisher (W-F) model for finite N and $v = 1/2$. What makes this case simple is that at each time step all individuals move. The probability that the time to the MRCA is t is given by (2). During t generations there are exactly $2t$ birth events in the ancestry of two individuals, and in the $v = 1/2$ case the phenotypic distance between two individuals follows a simple random walk with two steps in phenotype space per one time unit. Consequently, the distance between two siblings is always even. After some transient time the whole population will be constrained on the same sub-lattice of even, and then odd sites. The distance distribution of two individuals k and q, time t after their MRCA is

$$\Pr(X_k - X_q = x | T = t) = 2^{-2t} \binom{2t}{t + x/2} \tag{57}$$

where again x is always even. Consequently the probability $z(x)$ that two randomly chosen individuals are at distance x apart can be obtained from (6)

$$z(x) = \frac{1}{N-1} \sum_{t=1}^{\infty} \binom{2t}{t + x/2} \left(\frac{N-1}{4N} \right)^t \tag{58}$$

This sum can be evaluated using the identity

$$\sum_{t=1}^{\infty} \binom{2t}{t + x/2} \left(\frac{a}{4} \right)^t = \begin{cases} \frac{a}{\sqrt{1-a}(1+\sqrt{1-a})}, & x = 0 \\ \frac{a^{|x|/2}}{\sqrt{1-a}(1+\sqrt{1-a})^{|x|}}, & |x| \geq 2 \end{cases} \tag{59}$$

to obtain

$$z(x) = \begin{cases} \dfrac{1}{\sqrt{N}+1} & x = 0 \\ \dfrac{\sqrt{N}}{N-1} \left(\dfrac{N-1}{N + 2\sqrt{N} + 1} \right)^{|x|/2} & |x| \geq 2 \end{cases} \tag{60}$$

Hence, apart from the special $x = 0$ case, $z(x)$ decays exponentially in x. For fixed distances and $N \to \infty$ the asymptotic behavior is $z(x) = 1/\sqrt{N} + \mathcal{O}(1/N)$. The second moment of the distance distribution (60) is simply $2N$.

Now we turn to the strategies of the individuals. The strategies of the two players are the same if no mutations happened during time t to either player, which is the case with probability $(1-u)^{2t}$. Otherwise the two strategies are the same with probability $1/2$. Consequently, the conditional probability is

$$y(t) = (1-u)^{2t} + \frac{1}{2}[1 - (1-u)^{2t}] = \frac{1 + U^t}{2} \tag{61}$$

where we introduce the shorthand notations

$$U = (1-u)^2 \ , \quad M = N(1-U) + U \tag{62}$$

The probability y that two randomly chosen individuals have the same strategy becomes

$$y = \sum_{t=1}^{\infty} p(t) y(t) = \frac{1}{2} \left(1 + \frac{U}{M} \right) \tag{63}$$

where we have used (2) and (61).

Similarly, using (16) we obtain the probability g that two randomly chosen individuals have both the same strategy and the same phenotype

$$g = \frac{1}{2(\sqrt{N}+1)} + \frac{U}{2\sqrt{M}\left(\sqrt{N} + \sqrt{M}\right)} \tag{64}$$

These are exact results for arbitrary number of individuals N and mutation rate u. In the $N \to \infty$ and $u \to 0$ limit of the formulas (60), (63) and (64) with $\mu = 2Nu$ kept constant, we recover the $\nu \to \infty$ limits of the corresponding formulas (9), (15) and (19), apart from a factor two. This factor two is a peculiarity of the $v = 1/2$ case. Since here the distance between individuals is always even, there must be twice as many players at a given even distance. Note also that the variance of the cluster is 2ν both for $v = 1/2$ and for the continuous limit calculation.

For only two individuals, the general condition (48) simplifies to

$$\left(\frac{b}{c}\right)^*_{N=2} = \frac{1-y+z-g}{1-y-z+g} \tag{65}$$

which contains only quantities we have just calculated in this section. To obtain the exact $(b/c)^*$ for any other finite N we have to use the general expression (48), and obtain h analogously to (23) and using (20) and (21). The formulas for h and $(b/c)^*$ are too cumbersome to include here. We have, however, checked these formulas with computer simulations for many values of N. We explicitly simulated the W-F process and found the threshold $(b/c)^*$ value where the frequency of cooperators in the stationary state becomes larger than 1/2. Moreover, in the $N \to \infty$, $u \to 0$ limit with $\mu = 2Nu$ constant, we recover the continuous time formula (51).

4.2. **Excluding self interaction.** If cooperators cannot interact with themselves, we have

$$\begin{aligned} f_{C,i} &= 1 + \delta\left[b(m_i-1) - c(n_i-1)\right] \\ f_{D,i} &= 1 + \delta\left[bm_i\right]. \end{aligned} \tag{66}$$

Therefore the fitness of cooperators at position i becomes

$$w_{C,i} = 1 + \frac{\delta}{N}\left(b(m_i-1) - c(n_i-1) - \frac{b-c}{N}\sum_j m_j(n_j-1)\right) + \mathcal{O}(\delta^2) \tag{67}$$

which then leads to the expected change of cooperator frequency

$$\begin{aligned} \langle\Delta p\rangle =& \frac{\delta}{N^2}\Bigg[b\left\langle\sum_i m_i^2\right\rangle - c\left\langle\sum_i m_i n_i\right\rangle - \frac{b-c}{N}\left\langle\sum_{i,j} m_i m_j n_j\right\rangle \\ & - (b-c)\left\langle\sum_i m_i\right\rangle + \frac{b-c}{N}\left\langle\sum_{i,j} m_i m_j\right\rangle\Bigg] + \mathcal{O}(\delta^2). \end{aligned} \tag{68}$$

Two new correlation types in the neutral stationary state appear

$$\begin{aligned} \left\langle\sum_i m_i\right\rangle &= N\ \Pr(S_k = 1) = \frac{N}{2} \\ \left\langle\sum_{i,j} m_i m_j\right\rangle &= N^2\ \Pr(S_k = S_q = 1) = \frac{N^2}{2}y \end{aligned} \tag{69}$$

This then leads to the general expression analogous to (48) for the threshold ratio

$$\left(\frac{b}{c}\right)^* = \frac{(N-2)(z-h) + z - g}{(N-2)(g-h) - z + g} \tag{70}$$

The smallest valid population size is $N = 3$. In the $N \to \infty$ the threshold b/c ratio with self interaction (48) and without it (70) are the same (49) in the leading order, and their difference is only of order $1/N$.

4.3. **From averages to correlations.** Here we obtain the identities listed in (44). The variables m_i and n_i are fixed in any given state. Let us use the indicator function $\mathbb{1}$, which is $\mathbb{1}(A) = 1$ if event A is true and $\mathbb{1}(A) = 0$ if event A is false. Of course the stationary average of the indicator function is the stationary probability of an event

$$\langle \mathbb{1}(A) \rangle = \Pr(A) \tag{71}$$

and by $\mathbb{1}(A, B)$ we mean $\mathbb{1}(A \cap B) = \mathbb{1}(A)\mathbb{1}(B)$. Now in any given state we can express n_i and m_i by the indicator functions

$$\begin{aligned} n_i &= \sum_k \mathbb{1}(X_k = i) \\ m_i &= \sum_q \mathbb{1}(X_q = i)\mathbb{1}(S_q = 1). \end{aligned} \tag{72}$$

The sum in (44a) becomes

$$\sum_i m_i m_i = \sum_{k,q} \left[\mathbb{1}(S_k = 1)\mathbb{1}(S_q = 1) \sum_i \mathbb{1}(X_k = i)\mathbb{1}(X_q = i) \right] = \sum_{k,q} \mathbb{1}(S_k = S_q = 1)\mathbb{1}(X_k = X_q) \tag{73}$$

since the sum over i is simply

$$\sum_i \mathbb{1}(X_k = i)\mathbb{1}(X_q = i) = \sum_i \mathbb{1}(X_k = i,\ X_q = i) = \mathbb{1}(X_k = X_q). \tag{74}$$

Now taking the average of (73) in the stationary state we obtain

$$\left\langle \sum_i m_i^2 \right\rangle_0 = \sum_{k,q} \langle \mathbb{1}(S_k = S_q = 1,\ X_k = X_q) \rangle = \sum_{k,q} \Pr(S_k = S_q = 1,\ X_k = X_q), \tag{75}$$

where we have used identity (71). Since all individuals are equivalent in the stationary state, the above probabilities are the same for any pair of individuals, hence from now on we consider k and q as two randomly chosen individuals, and write

$$\left\langle \sum_i m_i^2 \right\rangle_0 = N^2 \Pr(S_k = S_q = 1,\ X_k = X_q). \tag{76}$$

The expression (44b) can be derived similarly, since

$$\sum_i m_i n_i = \sum_{k,q} \left[\mathbb{1}(S_q = 1) \sum_i \mathbb{1}(X_k = i)\mathbb{1}(X_q = i) \right] = \sum_{k,q} \mathbb{1}(S_q = 1)\mathbb{1}(X_k = X_q) \tag{77}$$

and taking the average of (77) in the stationary state leads to

$$\left\langle \sum_i m_i n_i \right\rangle_0 = \sum_{k,q} \Pr(S_q = 1, X_k = X_q) = N^2 \Pr(S_q = 1, X_k = X_q) \tag{78}$$

For the last expression (44c) we have

$$\begin{aligned}\sum_{i,j} m_i m_j n_j &= \sum_{k,q,l}\left[\sum_i \mathbb{1}(S_l=1, X_l=i)\right]\left[\sum_j \mathbb{1}(S_k=1, X_k=j)\mathbb{1}(X_q=j)\right] \\ &= \sum_{k,q,l}\mathbb{1}(S_l=1)\ \mathbb{1}(S_k=1,\ X_k=X_q)\end{aligned} \tag{79}$$

which in the stationary state becomes

$$\left\langle \sum_{i,j} m_i m_j n_j \right\rangle_0 = \sum_{k,q,l}\Pr(S_l=S_k=1,\ X_k=X_q) = N^3\ \Pr(S_l=S_k=1,\ X_k=X_q) \tag{80}$$

5. Outlook

5.1. **Moran dynamics.** In the Moran model we chose a random individual to die, and another (with replacement) to multiply with probability proportional to the player's payoff. The newborn then replaces the dead individual. Otherwise the dynamics is the same as in the W-F case. The behavior of the Moran model is also very similar to the W-F model, and the results can be written in an identical form in the $N \to \infty$ limit, by defining the appropriate variables.

We consider the neutral case of the Moran model first. Let us obtain the probability $\Pr(T=t)$ that the time to the most recent common ancestor (MRCA) of two randomly chosen individual is $T=t$. Let us calculate the probability P_{CA} that they had a common ancestor one update step before. It could happen only if the parent and the dying individuals were different, which happens with probability $1-1/N$. Then our two individuals have a common ancestor if one of them is the parent and the other is the newborn daughter, which has a probability $2\,\frac{1}{N}\,\frac{1}{N-1}$. Hence having a common ancestor in the previous update step is

$$P_{\text{CA}} = \left(1-\frac{1}{N}\right)\cdot 2\cdot\frac{1}{N}\cdot\frac{1}{N-1} = \frac{2}{N^2} \tag{81}$$

Consequently the probability that the MRCA is exactly time $T=t$ backward is

$$\Pr(T=t) = (1-P_{\text{CA}})^{t-1}P_{\text{CA}} = \left(1-\frac{2}{N^2}\right)^{t-1}\frac{2}{N^2} \tag{82}$$

If we introduce a rescaled time $\tau = t/(N^2/2)$, then in the $N\to\infty$ limit the coalescent time distribution (82) converges to the same density function (3) as we obtained for the W-F model.

Since in our model mutations (in strategies) and motion only happen at birth events, let us investigate the statistics of birth events in the Moran model. As we follow the ancestral lines of two randomly chosen individuals backward in time, we can obtain the probability P_{B} that a birth event happens in one update step, but the ancestral lines do not coalesce. In other words, P_{B} is the probability that at a given time one of the two individuals is the daughter but the other is not the parent. If the parent dies during this update step (which happens with probability $1/N$) one individual is the daughter with probability $2/N$ (and the other individual cannot be the parent). If the parent does not die (which happens with probability $1-1/N$) one of the individuals is the daughter and the other is not the parent with probability $2/N \cdot (N-2)/(N-1)$. Hence the probability that there is a birth event in the ancestry of either individual during one elementary time step is

$$P_{\mathrm{B}} = \left(1 - \frac{1}{N}\right) \cdot \frac{2}{N} \cdot \frac{N-2}{N-1} + \frac{1}{N} \cdot \frac{2}{N} = \frac{2(N-1)}{N^2} \tag{83}$$

In the continuous time limit with $\tau = t/(N^2/2)$, a birth event happens at rate N. Consequently a mutation happens at rate $\mu = Nu$ on the ancestral line of *two* individuals. Similarly, one of the two individual hops at rate $\nu = Nv$ in each direction. In other words the distance between the two individuals changes at rate ν in each direction. This means that the continuous time ($N \to \infty$) descriptions of the Moran and the W-F models are the same, but N must be used for the Moran and $2N$ for the W-F model in the definition of μ and ν. Hence all $N \to \infty$ results of Section 2 are also valid for the Moran model. (Note that the diffusion coefficient of the cluster is $D = v/N$.)

All formulas of Section 3 are almost identical to those for W-F model. The average frequency of cooperators depends on the change of cooperators very similarly to (32)

$$\langle p \rangle = \frac{1}{2} + N\frac{1-u}{u}\langle \Delta p \rangle_{\mathrm{sel}} \tag{84}$$

Instead of the fitness of the W-F model (28), we have a very similar expression for the fitness after one elementary step

$$w_{C,i} = \frac{N-1}{N} + \frac{f_{C,i}}{\sum_j [m_j f_{C,j} + (n_j - m_j) f_{D,j}]} \tag{85}$$

where the payoffs are again given by (27). Here the first term corresponds to the cooperator staying alive, and to second to it being chosen for reproduction. In the $\delta \to 0$ limit (85) becomes

$$w_{C,i} = 1 + \frac{\delta}{N}\left(bm_i - cn_i - \frac{b-c}{N}\sum_j m_j n_j\right) + \mathcal{O}(\delta^2) \tag{86}$$

Note that this is exactly the fitness of the W-F process (30) with a scaled selection strength $\delta' = \delta/N$. Hence all results of Section 3, and in particular the citical b/c ratio (50) are also valid for the Moran model.

5.2. **General payoff matrix.** Instead of the payoff matrix (26) of the simplified Prisoner's Dilemma (PD) game, we study now a general payoff matrix

$$\begin{pmatrix} R & S \\ T & P \end{pmatrix} \tag{87}$$

A similar derivation to the one presented in Section 3 leads to the condition for cooperation

$$(R-S)g + (S-P)z > (R-S-T+P)\eta + (S+T-2P)h \tag{88}$$

in the $N \to \infty$ limit, which is the analogous formula to (49). Here a new type of three point correlation must be introduced

$$\eta = \Pr(S_l = S_k = S_q,\ X_k = X_q) \tag{89}$$

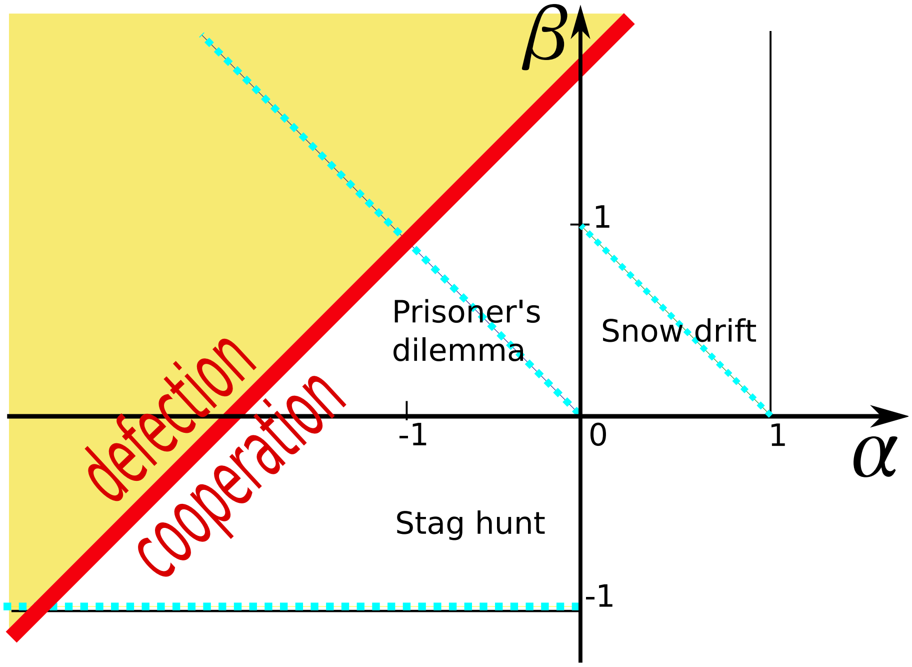


FIGURE 2. "Snow drift", "Stag hunt" and "Prisoner's dilemma" games correspond to three distinct regions in the (α, β) plane, bounded by black lines. The red (thick) line (94) marks the boundary between defection (yellow-shaded) and cooperation (white). The blue (thicker dashed) lines depict the corresponding simplified payoff matrices.

In the $\nu \to \infty$ and $\mu \to 0$ limit the correlations are

$$z = \frac{1}{2\sqrt{\nu}} \qquad g = \frac{1}{2\sqrt{\nu}}\left(1 - \frac{\mu}{4}\right)$$
$$h = \frac{1}{2\sqrt{\nu}}\left(1 - \mu\frac{2+\sqrt{3}}{8}\right) \qquad \eta = \frac{1}{2\sqrt{\nu}}\left(1 - \mu\frac{3+\sqrt{3}}{8}\right) \tag{90}$$

up to $\mathcal{O}(1/\nu)$ and $\mathcal{O}(\mu^2)$ terms. Here z, g, and h were obtained as limits of the general expressions (10), (19), and (24) respectively. The value of η was derived analogously to (23). By substituting these correlations into (88) we finally arrive at the general condition for cooperation

$$T - S < (R - P)(1 + \sqrt{3}) \tag{91}$$

For the simplified PD game (26) we recover (52) in the leading order.

For a non-degenerate payoff matrix, with the exchange of players $R > P$ can always be achieved. Then under weak selection one can define an equivalent matrix

$$\begin{pmatrix} 1 & \alpha \\ 1+\beta & 0 \end{pmatrix} \tag{92}$$

with only two parameters

$$\alpha = \frac{S - P}{R - P}, \quad \beta = \frac{T - R}{R - P} \tag{93}$$

In these variables the condition for cooperation (91) becomes

$$\beta < \alpha + \sqrt{3} \tag{94}$$

which describes a straight threshold line in the (α, β) plane (see Figure 2).

In Figure 2 we show how this threshold line (94) divides the (α, β) plane into a cooperative and a defective half plane. Three regions, bounded by black lines, correspond to the "Snow drift", the "Stag hunt" and the "Prisoner's dilemma" games. The blue straight lines on the (α, β) plane correspond to the following representative simplified payoff matrixes

$$\begin{array}{lll} \text{Snow drift} & \begin{pmatrix} b-c/2 & b-c \\ b & 0 \end{pmatrix} & \beta = 1-\alpha, \text{ with } 0<\alpha<1 \\ \text{Stag hunt} & \begin{pmatrix} b-c & -c \\ 0 & 0 \end{pmatrix} & \beta = -1, \text{ with } \alpha<0 \\ \text{Prisoner's dilemma} & \begin{pmatrix} b-c & -c \\ b & 0 \end{pmatrix} & \beta = -\alpha, \text{ with } \alpha<0 \end{array} \tag{95}$$

Form the general condition (91) we can deduce the condition for cooperation for these simplified games. There is always cooperation in the simplified Snow drift game. Cooperation is favored in the simplified Stag hunt game only for $b/c > 1 + 1/(1+\sqrt{3})$. In the simplified PD game cooperators win for $b/c > 1 + 2/\sqrt{3}$ in agreement with (52).

5.3. **Randomly changing phenotypes.** Here we replace the one-dimensional phenotype space with an infinite-dimensional phenotype space. We do not model the number of dimensions explicitly, but simply assume that every mutation causes a jump to a new unique phenotype. Now the only way that two individuals can have the same phenotype is if there are no phenotypic mutations in their ancestry back to the time of their most recent common ancestor. This property is called *identity by descent* in population genetics and this mutation model known as the infinitely-many-alleles, or simply infinite-alleles, mutation model [6, 7].

Let $\tilde{v}$ be the probability that the phenotype of an offspring differs from that of its parent. Note that in the one-dimensional model, there is a mutation probability of v in each direction. As before, in the limiting ($N \to \infty$) model with time rescaled appropriately, the phenotypic mutation rate to two individuals is equal to ν. In the Wright-Fisher model we have $2N\tilde{v} \to \nu$ (and $N\tilde{v} \to \nu$ in the Moran model), where the arrows correspond to the limit $N \to \infty$. The definition of $\mu = 2Nu$ in the Wright-Fisher model ($\mu = Nu$ in the Moran model) is the same as before.

Given a coalescence time τ between a pair of individuals,

$$\zeta(\tau) = e^{-\nu\tau} \tag{96}$$

is the probability that they have the same phenotype. Therefore, in the $N \to \infty$ limit, the correlations defined in (1) become

$$\begin{aligned} z &= \frac{1}{1+\nu} \\ g &= \frac{1}{2}\left(\frac{1}{1+\nu} + \frac{1}{1+\mu+\nu}\right) \\ h &= \frac{1}{2}\left[\frac{1}{1+\nu} + \frac{1}{3+\mu+\nu}\left(\frac{1}{1+\nu} + \frac{1}{1+\mu} + \frac{1}{1+\mu+\nu}\right)\right] \end{aligned} \tag{97}$$

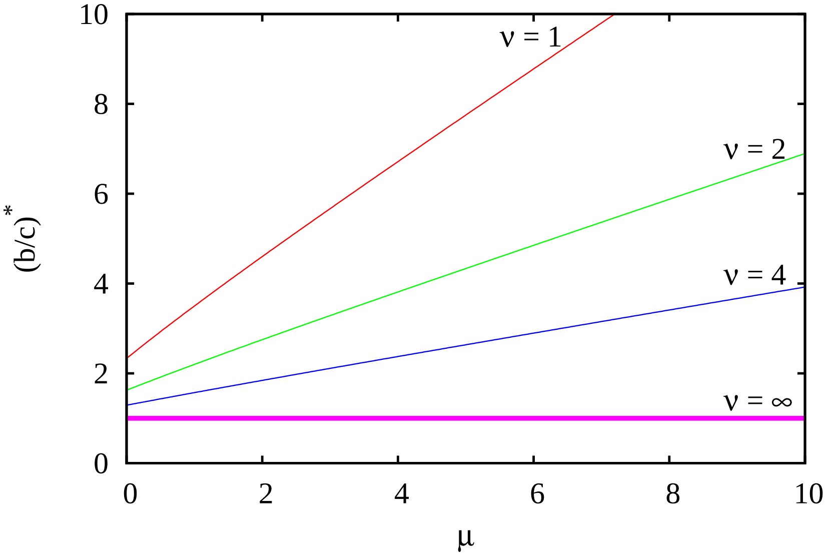


FIGURE 3. Exact threshold b/c ratio (98) for randomly changing phenotypes for $N \to \infty$. Cooperation is most favored in the $\nu \to \infty$ limit, where $(b/c)^* = 1$. Note that the lines for finite values of ν are not straight.

The calculation goes analogously to that of Section 2. The threshold parameters (49) for cooperation to be favored becomes

$$\left(\frac{b}{c}\right)^* = \frac{\nu(3 + 2\mu + \nu) + (1 + \mu)(3 + \mu)}{\nu(2 + \mu + \nu)} \tag{98}$$

This is plotted in Figure 3, which can be compared to the corresponding Figure 1 for the one-dimensional model.

Cooperation is most favored when ν is large because in this case two individuals that share the same phenotype will almost surely have the same strategy. We have

$$\left(\frac{b}{c}\right)^* = 1 + \frac{1 + \mu}{\nu} + O(\nu^{-2}) \tag{99}$$

In the $\nu \to \infty$ limit, $(b/c)^* = 1$, *i.e.* cooperation is favored whenever the benefit b from cooperation is larger than the cost c.

For general payoff matrices (87), we restrict our calculation to the $\mu \to 0$ limit. The calculation is completely analogous to that of Section 5.2. First we calculate the three point correlation η, which is defined in (89). Up to first order in μ we obtain

$$\eta = \frac{1}{1 + \nu}\left[1 - \mu\frac{9 + 7\nu + 2\nu^2}{4(1 + \nu)(3 + \nu)}\right] \tag{100}$$

Substituting this expression together with (97) into the general condition (88) for cooperation, we finally obtain

$$T - S < (R - P)\,\frac{(1 + \nu)(3 + 2\nu)}{3 + \nu} \tag{101}$$

This result is valid for general values of ν. For $\nu \to 0$ condition (101) becomes $T - S < R - P$, while in the $\nu \to \infty$ limit it is simply $R > P$.

By using the scaled variables α, β, introduced in (92), condition (101) is again a straight line in the (α, β) plane. For $\nu \to 0$ there is no cooperation in the PD region (see this region in Figure 2), but for $\nu \to \infty$ the whole plane corresponds to cooperation.

References

[1] P. A. P. Moran (1975) Wandering distributions and electrophoretic profile. *Theor. Popul. Biol.* **8**:318-330.

[2] J. F. C. Kingman (1976) Coherent random-walks arising in some genetic models, *Proc. R. Soc. Lond. A* **351**:19-31.

[3] N. G. van Kampen (1997) Stochastic Processes in Physics and Chemistry. 2^{nd} ed., North-Holland, Amsterdam.

[4] S. Redner (2001) A Guide to First-Passage Processes, Cambridge University Press, New York.

[5] I. S. Gradshteyn and I. M. Ryzhik (2007) Table of Integrals, Series, and Products. 7^{nd} ed., Elsevier, Amsterdam.

[6] G. Malécot (1946), C. R. Acad. Sci., **222**, 841-843.

[7] M. Kimura and J. F. Crow (1964), Genetics **49**, 725-738.

Sympathy and similarity: The evolutionary dynamics of cooperation

Karl Sigmund[1]

Faculty of Mathematics, University of Vienna, 1090 Vienna, Austria; and International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria

The advantage of mutual help is threatened by defectors, who exploit the benefits provided by others without providing benefits in return. Cooperation can only be sustained if it is preferentially channeled toward cooperators and away from defectors. But how? A deceptively simple idea is to distinguish cooperators from defectors by tagging them. It clearly is in the interest of cooperators to use some distinctive cue to assort with their like. Such an assortment, however, conflicts with the interests of the cheaters, who have every incentive to also acquire that tag. This makes for an inherently unstable situation. The history of evolutionary thinking on this issue is long. An article in this issue of PNAS by Antal et al. (1) opens new ground by providing an in-depth analysis of a selection-mutation model.

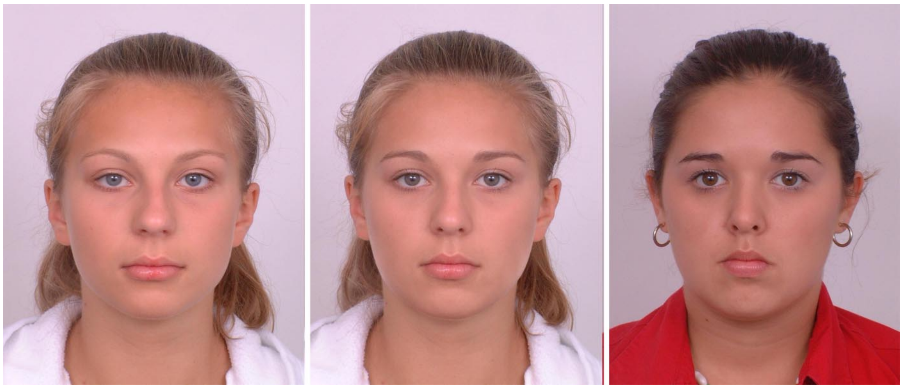


Fig. 1. Face transitions. Players in a game theoretic experiment are provided with pictures of their partners who, through digital sorcery, are made to look like themselves, to a greater or lesser extent. Here, the face in the middle is the result of a 60:40 mix of the other 2 faces. Players preferentially trust coplayers who look more like themselves. Thus, familiarity enhances trust. With permission from Lisa DeBruine, see ref. 14.

The first to investigate a tag for altruism was W. D. Hamilton (2). He conceived what he called a supergene, able to produce (*i*) a distinctive phenotypic trait, (*ii*) the faculty to recognize the trait in others, and (*iii*) the propensity to direct benefits toward bearers of that trait, even though this entails a fitness cost. Soon afterward, Richard Dawkins described Hamilton's thought experiment by using as phenotypic trait the fanciful example of a green beard. The supergene was now termed "green beard gene," in part to acknowledge its inherent unlikelihood. "Too good to be true," were Dawkins' words (3): for the gene would have to be able to program for 3 effects, namely the feature, its recognition, and the altruistic propensity.

The green-beard concept relates to both major approaches to cooperation in evolutionary biology, namely kin selection (2) and reciprocal altruism (4). It helps in promoting assortment between cooperators; as a result, cooperators can get more than they give, so that altruism becomes a thriving business. Because wearers of green beards both confer and receive benefits, the tag works as a kind of promise that the altruistic action will be returned, not necessarily by the recipient, but by another member of the green-bearded guild. In this sense, the green beard mediates an indirect form of reciprocation, through third parties. In the usual models of indirect reciprocity, "good guys" are recognized by their reputation, which is based on their past deeds (5). Here, however, recognition is ensured by a phenotypic trait, which is a less sophisticated (and possibly less reliable) signal.

Mostly, the green beard is studied in the context of kin selection. If you carry a green beard, your relatives are likely to carry one, too. Directing benefits at green-bearded individuals confers the benefits preferentially to your kin and raises your indirect fitness (because your kin shares your genes with a higher-than-average probability). In many cases, kin are living close by. But the viscosity of the population (to use another term by Hamilton) is not enough to guarantee a local increase in cooperation, because it is counterbalanced by a local increase in competition. Limited dispersal alone is therefore not enough. A gene for kin recognition can help to direct positive rather than negative effects toward relatives. But it is important to realize that the green beard can promote altruism beyond the realm of the family.

Some 10 years ago, it was found that green beards are not as implausible as their name suggests. In particular, Haig (6) remarked that genes for homophilic cell adhesion could perform all 3 tasks required from a green-beard gene (trait, recognition, and action) by coding for a surface protein that allows them to stick to copies of themselves on other cells. A few years later, it was found that csA genes in *Dictyostelium discoideum* fit the bill (7). In hard times, these amoeba literally stick together to form stalks for dispersing their spores. A similar gene has also been discovered in flocculating yeast cells (8). Other candidates for more sophisticated green-beard effects have been found in ants and lizards.

An obvious way to cheat is to grow a green beard but skip the altruism. For homophilic cell adhesion, this seems barely feasible. In other examples, cheating may be prevented by genetic constraints. But in principle, one would expect that a tight link between a gene for altruistic behavior and a gene for tag recognition will ultimately be broken, and cooperation be destroyed. Surprisingly, it turned out that if the link is not too tight (but not too loose either), a dynamic regime of cooperation can emerge, based on tag diversity. Whenever some tag becomes too frequent, it can be faked by defectors, but cooperative behavior subsists nevertheless, by allying itself with another tag. This phenomenon has been termed "beard chromodynamics," to suggest that green beards can over time be replaced by red,

Author contributions: K.S. wrote the paper.

The author declares no conflict of interest.

See companion article on page 8597.

[1] E-mail: karl.sigmund@univie.ac.at.

or blue, or yellow beards as rallying signals for cooperators (9, 10).

The underlying principle is that of a shibboleth, or secret handshake. But such a specially-contrived trait, evolved for the purpose of signaling cooperation, is not always necessary.

Tag-based cooperation can also rely on self-similarity. All that is needed is some general means to recognize what is like yourself and what is not, i.e., to distinguish "us" from "them."

With familiars, you need no badge, or password. This has been called the "armpit effect" (by Dawkins see ref. 3). Although an obvious variation of the green-beard principle could mediate, in principle, symbiosis between 2 different species, the armpit effect is self-referential. You need not sprout a special recognition device but simply check whether the other looks, smells, or sounds like you.

Mechanisms based on self-similarity are commonly used among cells of an organism or among members of a species. Kin recognition seems widespread: it is useful, not only for promoting nepotism, but also for avoiding incest (11). Bats or birds recognize their offspring on crowded cave roofs and cliff faces through vocalizations; hamsters and wasps pick up the odor of their nest or colony, etc. Interestingly, these faculties seem always acquired through imprinting, rather than genetically encoded. Thus, they indicate in-group rather than kin. This use of associative learning is well supported by theory (12).

An armpit effect has been recently found in hamsters (13). Self-similarity appears to work in humans, too: we like our like. Neat economic experiments show that players preferentially trust similar-looking coplayers (14) (Fig. 1). (The players are provided with pictures of their ostensible partners, and these photos are manipulated to look to a greater or lesser degree like themselves). Clearly, such cues for self-similarity can be enhanced by cultural means. Many groups provide their members with characteristic uniforms, badges, tattoos, ties, haircuts, hangouts, accents, musical tastes, or slang idioms.

In most tag-based models, the tags are discrete; you either look like me or you do not. In general, defectors can be overcome only for a restricted range of recombination between tag and behavior (cf. refs. 15–17). However, similarity is likely to be a question of degree; you can look more or less like me. In the case of continuous graduation, it is likely that cooperative behavior is addressed toward all those who are tolerably similar.

Economic experiments show that players preferentially trust similar-looking coplayers.

Such models show intriguing patterns: cliques of similar cooperators grow, are beset and undermined by defectors, and regroup around other phenotypes (18, 19). Extending tolerance to a larger range of tag values enlarges the basis of collaboration, whereas restricting tolerance shields from exploiters: this leads to endlessly fluctuating "tides of tolerance" (20).

In the model of Antal et al. (1), members of a well-mixed population of constant size N are distinguished by a tag that can take infinitely many values and is coded by integers.

Defectors help nobody, and cooperators provide help exclusively to members of their own tag group. From time to time, individuals produce offspring in numbers proportional to their fitness. Some N of these offspring are randomly chosen to form the next generation. Offspring inherit from their parent both their behavior (cooperator or defector) and their tag, up to mutation. Each configuration of the population is specified by the number of defectors and cooperators for each tag. The expected payoff values for defectors and cooperators can easily be computed in terms of conditional probabilities (e.g., for defectors to interact with cooperators, etc.). This specifies the configurations for which cooperators are sufficiently assorted with other cooperators to earn more than defectors do. But the configurations move and cluster in a very fluid manner through the range of possible tags. It needs considerable mathematical dexterity to average the payoffs over all configurations in the stationary state. This yields, under the limiting assumption of weak selection, a condition for cooperators to be more frequent than defectors in the long term, requiring that the benefit-to-cost ratio exceeds a specific threshold. Under the most favorable conditions, i.e., when mutations between tags are frequent and mutations in the behavior rare, that threshold is slightly larger than 2. In contrast to previous models (9, 15, 16), no additional requirements on spatial population distribution are used. The analysis of several limiting cases shows that the results depend significantly on mutation structure, about which empirical data are lacking at present. The elusive nature of the game of hide and seek between cooperators and defectors, an age-long spur for biological and cultural evolution, continues to challenge experimentalists and theoreticians alike.

1. Antal T, et al. (2009) Evolution of cooperation by phenotypic similarity. *Proc Natl Acad Sci USA* 106:8597–8600.
2. Hamilton WD (1964) The genetical basis of social behavior I. *J Theor Biol* 7:1–16.
3. Dawkins R (1982) *The Extended Phenotype* (Oxford Univ Press, Oxford).
4. Trivers R (1971) The evolution of reciprocal altruism. *Q Rev Biol* 46:35–57.
5. Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. *Nature* 437:1292–1298.
6. Haig D (1996) Gestational drive and the green-beard placenta. *Proc Natl Acad Sci USA* 93:6547–6551.
7. Queller DC, et al. (2003) Single-gene green beard effects in the social amoeba *Dictyostelium discoideum*. *Science* 299:105–106.
8. Smukalla S, et al. (2006) Flo1 is a variable green-beard gene that drives biofilm-like cooperation in budding yeast. *Cell* 135:726–737.
9. Jansen VAA, van Baalen M (2006) Altruism through beard chromodynamics. *Nature* 440:663–664.
10. Traulsen A, Nowak MA (2007) Chromodynamics of cooperation in finite populations. *PLoS One* 2:e270.
11. Pfennig DW (2002) Kin recognition. *Encyclopedia of Evolution*, ed Pagel M (Oxford Univ Press, Oxford), pp 592–596.
12. Lehmann L, Perrin N (2002) Altruism, dispersal, and phenotype-matching kin recognition. *Am Nat* 159:451–467.
13. Mateo JM, Johnston RE (2000) Kin recognition and the ''armpit effect'': Evidence of self-referent phenotype matching, *Proc R Soc London Ser B* 267:695–700.
14. Krupp DB, DeBruine LM, Barclay P (2008) A cue of kinship promotes cooperation for the public good. *Evol Hum Behav* 29:49–55.
15. Rousset R, Roze D (2007) Constraints on the origin and maintenance of genetic kin recognition. *Evolution* 61:2320–2330.
16. Axelrod R, Hammond RA, Grafen A (2004) Altruism via kin-selection strategies that rely on arbitrary tags with which they coevolve. *Evolution* 58:1833–1838.
17. Gardner A, West SA (2007) Social evolution: The decline and fall of genetic kin recognition. *Curr Biol* 17:R810–R812.
18. Riolo RL, Cohen MD, Axelrod R (2001) Evolution of cooperation without reciprocity. *Nature* 414:441–443.
19. Traulsen A, Schuster HG (2003) Minimal model for tag-based cooperation. *Phys Rev E* 68:046129.
20. Sigmund K, Nowak MA (2001) Tides of tolerance. *Nature* 414:403–404.