Common Interest and Signaling Games: A Dynamic Analysis

Manolo Martínez and Peter Godfrey-Smith

Abstract

We present a dynamic model of the evolution of communication in a Lewis signaling game while systematically varying the degree of common interest between sender and receiver. We show that the level of common interest between sender and receiver is strongly predictive of the amount of information transferred between them. We also discuss a set of rare but interesting cases in which common interest is almost entirely absent, yet substantial information transfer persists in a *cheap talk* regime, and offer a diagnosis of how this may arise.

1 Introduction

An important recent development in the naturalistic study of communication is the systematic investigation of simple, computationally tractable models. Such models are severely idealized in many respects, but there are invaluable gains in the explicitness and rigor of the results obtained by working on them, alongside the familiar approach of engaging in informal discussion of more realistic examples. Some areas of concern to philosophers that this research program has already shed light on are the difference between assertions (indicatives) and directives (imperatives) (Huttegger 2007; Zollman 2011); signaling in social dilemmas (Wagner 2014); deception (Zollman, Bergstrom and Huttegger 2013; Martínez 2015), and vagueness (O'Connor 2014).

Formulating the problem of communication in a way that makes it ammenable to a rigorous treatment of this kind is in itself a major philosophical contribution. Most of the work cited above is based on Lewis's (1969/2002) model of signaling. In this model a *sender* (or, for Lewis, 'communicator') sends messages to a *receiver*, (or, for Lewis, an 'audience') and both parties receive a payoff that depends on the state the world is in when the message is sent and the act performed by the receiver in response. (In this paper we focus on so-called *cheap talk* games, in which payoffs do not depend on the type of message sent.) The message sent by the sender on a given occasion is decided by a *sender's strategy*: a function that takes states (that is, members of a set S of mutually exclusive and jointly exhaustive ways the world can be) to a probability distribution over the set M of possible messages. The act performed by the receiver is decided by a *receiver's strategy*: a function that takes each member of M to a probability distribution over the set A of possible acts. A *signaling game* is individuated by two *payoff matrices* that give the payoffs for sender and receiver for each combination of state and act, together with a distribution that gives the unconditional probabilities of states. A *sender-receiver configuration* is individuated by a signaling game, a sender's strategy and a receiver's strategy.

So, for example, a certain signaling game, SG, is univocally described by giving, first, the payoff matrices in Table 1 and, second, the distribution for S $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ – that is, by stating that the three states the world can be in are equiprobable. And, for example, a sender-receiver configuration is individuated by SG together with the sender's and receiver's strategies in Table 2.

That is, the sender will always send M_1 in S_1, and M_2 in S_2, and will throw a biased coin in S_3, so as to send M_1 with a probability of two thirds, M_3 with a probability of one third. The receiver's strategy can be read analogously.

In Lewis's original discussion, sender and receiver are rational agents with complex intentional profiles, and emphasis is put on sender-receiver configurations that achieve various kinds of equilibrium states. In Skyrms's (1996, 2010) groundbreaking reinterpretation of the Lewisian framework, in contrast,

	S_1	S_2	S_3
A_1	5,0	2,4	0,5
A_2	6,5	0,0	1,6
A_3	0,6	6,6	5,3

Table 1: Two payoff matrices. The pair of numbers in each cell represent, respectively, the sender's and the receiver's payoffs for a given action (A) performed by the receiver in a given state of the world (S). The payoff matrix for the sender cam be reconstructed by taking the first member of the pair of numbers in each cell; that for the receiver, by taking the second member.

	S_1	S_2	S_3
M_1	1	0	$\frac{2}{3}$
M_2	0	1	0
M_3	0	0	$\frac{1}{3}$

	M_1	M_2	M_3
A_1	1	$\frac{1}{2}$	0
A_2	0	$\frac{1}{2}$	0
A_3	0	0	1

Table 2: A sender's and a receiver's strategies

what counts is not the players's rational appreciation of the payoff situation, but the way in which various selection processes (evolution, reinforcement learning, and imitation) can shape the strategy of agents, who may be individually very unintelligent, as a result of those strategies being more or less successful in securing payoffs.

In the present paper, we apply these methods to some long-standing questions about the relationship between communication and common interest. Many theorists, both in philosophy and other fields, have seen communication as a fundamentally cooperative affair, an interaction between agents whose interests are at least fairly well aligned. This has been a common theme across a range of literatures, including speech act theory (Grice 1957), Ruth Millikan's naturalistic theory of intentionality (1984), and a range of recent work on the evolution of human behavior (Tomasello 2008; Sterelny 2012). Lewis himself assumed that common interest "predominates" in his original model (Lewis 1969/2002, 10). In the context of the Lewis model, there is common interest if sender and receiver tend to want the same acts performed in a given state of the world. Interests are divergent when the two agents want different pairings of actions and states. An intuition that many have shared is that if the interests of sender and receiver are too divergent, then a receiver will be unwise to trust anything a sender says. If receivers stop listening, there is no point in talking. So divergence of interests should eventually make communication collapse. This intuition is fundamentally a dynamic one; it predicts that when sender and receiver diverge enough in interests, a particular outcome should occur. How, then, do these ideas fare when cast in a formal model of behavioral change, using a dynamic version of the Lewis model? What role does common interest have in producing and maintaining communicative interactions? Those are the themes of this paper.

Earlier work has already shed some light on this question. Skyrms (2010) considered a few cases of imperfect alignment of interests in a Lewis signaling model, and showed that communication could be an equilibrium state in these cases. Skyrms discussed just a handful of cases, though. A classic model from economics, Crawford and Sobel (1982), gives a more general and rigorous treatment of the consequences of divergent interests for a static model that has both similarities and differences from the Lewis set-up. They imagined a situation in which a sender wants to exaggerate their quality (or another relevantly similar state of the world), to some degree, and the receiver wants not to be taken in by the exaggeration. Crawford and Sobel found that as interests diverge more and more, fewer distinct signals can be used at equilibrium, until signaling collapses into a "pooling" outcome in which the sender makes no distinction between different states of the world. This result is in accordance with the intuition, outlined above, that when interests diverge too far, senders will say nothing worth listening to.

In both economics and biology, a rich set of models has been developed that explore the consequences of differential signal cost in enforcing honesty when interests diverge. When signaling itself is a costly action, there are situations in which it is plausible that only honest senders can afford to send a signal of a given kind. The first model to explore this idea was offered in economics by Spence (1973), and applied to the case of job markets. Zahavi, independently, soon after applied the same principle to biology (1975), where the choice of mates by females replaced the choice of employees (see also Grafen 1990a, b; Maynard-Smith and Harper 2003; and Zollman, Bergstrom and Huttegger 2013). Since then, a wide range of models of this kind have been developed. In some models, the costs need only be operative when the population is not at equilibrium (Lachmann, Szamado and Bergstrom 2001).

The detailed development of costly signaling models may have fostered the impression that communication is very difficult to maintain in situations where signal costs are entirely absent and interests do not align. Some well-known games do give this impression. However, this impression is somewhat misleading. In earlier work of our own, (Godfrey-Smith and Martínez 2013), we used computerized search methods to assess the value of some exact measures of common interest as predictors of the viability of communication in a cheap-talk Lewis model. By looking far outside the set of familiar games, we found that communication could persist in some situations characterized by extremely low levels of common interest. We also found a general predictive relationship between our measures of common interest and the viability of communication. This earlier work, however, focused entirely on the existence of Nash equilibria[1] and contained no dynamical models. It did not investi-

[1]In this context, a Nash equilibrium is a sender-receiver configuration in which neither sender nor receiver can increase their expected payoff by unilaterally changing their strategy.

gate how accessible to evolution equilibria were.[2] The present paper considers the relationship between communication and common interest using dynamic methods. We ask how different degrees of common interest affect the evolutionary trajectories of populations of senders and receivers interacting in accordance with a Lewis model.

Section 2 describes the model used: a family of Lewis sender-receiver games, embedded in an evolutionary model on which quantitative measures of communication and common interest are defined. Section 3 discusses the main results regarding the relation between the degree of common interest present in a game and the maintenance of communication, while Section 4 takes a closer look at the case of very low common interest. Section 5 assesses the significance of our results, and offers conclusions.

2 The Model

Our model uses a Lewis sender-receiver game of the kind characterized in the introduction, and then embeds this game in an evolutionary model in which change is described with the replicator dynamics (Hofbauer and Sigmund 1998, chap. 7; Sandholm 2010, 126). Specifically, consider a signaling game, a set of possible sender's pure strategies $\mathfrak{S} = \{\sigma_1, \ldots, \sigma_q\}$, and a set of possible receiver's pure strategies $\mathfrak{R} = \{\varrho_1, \ldots, \varrho_r\}$. Instead of a single sender and a single receiver, we have a population of senders, and another of receivers. The sender population can be characterized in terms of q behavioral types, each one of them implementing a different strategy in $\mathfrak{S}$, and their associated frequencies. The receiver population is, similarly, characterized by r types and their associated frequencies. The frequencies of the different sender and receiver types are $X = \{x_1, \ldots, x_q\}$ and $Y = \{y_1, \ldots, y_r\}$.

Members of the sender population are assumed to interact with members of the receiver population. The average payoff for the sender type that follows strategy σ_i when dealing with a receiver following strategy ϱ_j is π^{σ}_{ij}. The receiver in that encounter gains π^{ϱ}_{ji}. These payoffs are easily calculated from the payoff matrices, the players' strategies, and the unconditional probabilities of states.

The *average payoff* for a sender type is the weighted average of the payoffs this type gets with each type present in the receiver population: $\overline{\pi}^{\sigma}_i = \sum_j \pi^{\sigma}_{ij} \cdot y_j$. Mutatis mutandis for the receiver: $\overline{\pi}^{\varrho}_i = \sum_j \pi^{\varrho}_{ij} \cdot x_j$.

[2] Wagner (2012), discussed below, and Wagner (2014) also use dynamic methods to study the emergence of communication in situations of significant conflict of interest.

Finally, the average payoff for the entire sender population is the weighted average of the averages per type: $\overline{\pi}^{\sigma} = \sum_i x_i \overline{\pi}^{\sigma}_i$. Mutatis mutandis for the receiver: $\overline{\pi}^{\varrho} = \sum_i y_i \overline{\pi}^{\varrho}_i$.

If sender and receiver populations follow the *two-population replicator dynamics*, the rate of change over time of the frequency of each type is given by the following differential equations:

$$\dot{x}_i = x_i \cdot (\overline{\pi}^{\sigma}_i - \overline{\pi}^{\sigma}) \tag{1}$$

$$\dot{y}_i = y_i \cdot (\overline{\pi}^{\varrho}_i - \overline{\pi}^{\varrho}) \tag{2}$$

We have now embedded a Lewis model within an evolutionary framework. Our next topic is the characterization of common interest between sender and receiver.

As in Godfrey-Smith and Martínez (2013), we use C as a measure of common interest between sender and receiver. C formalizes the following idea: sender and receiver see perfectly eye to eye insofar as the outcome they most prefer coincides in every state, their second preference coincides too, and so on down to the least preferred outcome. Their interests diverge gradually as these preference rankings diverge.

C is calculated as follows. For each state (i.e., each column in the sender and receiver payoff matrices), we calculate the *Kendall tau distance* (the number of pairwise disagreements in the ranking of acts) between sender and receiver payoffs. For example, in the payoff matrix in Table 1, the Kendall tau distance for state S_1, τ_{S_1}, is 2: sender and receiver disagree about which member is preferable in the pairs of acts (A_1, A_3) and (A_2, A_3), but agree that A_2 is preferable to A_1. τ_{S_2} is 0: they agree completely on the preference ranking for acts in that state. τ_{S_3} is also 2. An average distance is then calculated, using the unconditional probabilities of states as weights:

$$\tau = \sum_i \Pr(S_i) \cdot \tau_{S_i}$$

Finally, τ is rescaled so as to have 0 as no common interest, and 1 as perfect common interest. For n states this yields

$$C = 1 - \frac{2}{n(n-1)} \cdot \tau$$

C is a very coarse-grained measure of common interest.[3] In games with 3 equiprobable states and 3 acts, there are only 10 possible values of C. How-

[3]For example, it does not give any special role or weighting to actions that yield the *best* payoff for sender and receiver. There is some reason to

ever, as we will see, it is strongly predictive of the possibility of communication.

Next we consider how to describe communication itself in such a set-up. We say that a sender-receiver configuration contains informative signaling when the signals sent carry some information about the state of the world, and the acts performed carry information about the signal sent. These relationships are measured as *mutual information*. This is a widely-used concept, originally due to Shannon (see Shannon and Weaver 1949), that measures the amount of association between two variables, the extent to which the value of one predicts the value of the other. Mutual information is symmetrical and its value ranges between a minimum of 0 (no association) and a maximum dependent on the amount of entropy (uncertainty) in the two variables. It is calculated as follows:

The (unconditional) entropy of states is given by

$$H(S) = -\sum_i \Pr(S_i) \log_2(\Pr(S_i))$$

And the entropy of states conditional on acts is given by

$$H(S|A) = -\sum_i \Pr(A_i) H(S|A = A_i)$$

where $\Pr(S_i)$ is the unconditional probability of state S_i. Finally, the mutual information between states and acts is given by

$$I(S; A) = H(S) - H(S|A)$$

In games with 3 states, 3 messages, and 3 acts, if $I(S; A) = \log_2 3$, the sender's strategy is a bijection between S and M, and the receiver's strategy a bijection between M and A. We will refer to configurations in which both sender's and receiver's strategies have this property as *signaling systems* – we take this notion from Lewis (1969/2002), but our use differs from his in that we are placing no constraints on the payoffs received by sender and receiver, while for Lewis players engaging in a signaling system always obtain maximum payoffs. If $I(S; A) = 0$ nothing whatsoever can be said about the state of the world from the act the receiver performs – this corresponds to the absence of communication.

think that a match in these actions, for a given state, should be particularly important in maintaining communication. However, when we experimented with a weighting of this kind, the result was a less predictively useful measure than C.

Next we describe the relationships between the states of populations of senders and receivers, on one hand, and the measure of communication outlined above. As we have set up our model, at any given time a range of different types may exist in each of the two populations – the sender population and receiver population. Thus, a great range of different communicative interactions are assumed to be taking place – there is not a single sender-receiver configuration present, in the sense we introduced earlier. However, we aim to give a general characterization of the sender-receiver relationships that exist at each time. We do this in a way that has become common in models of this kind; we "translate" the pair of population structures that are present at a time into a single sender-receiver configuration by averaging over the different individual-to-individual interactions that are possible given the state of the two populations.[4] For each state of the world, there is a probability distribution over messages that is determined by the state of the sender population. Thus there is a population-wide "sender's strategy" instantiated for that state. Similarly, for each available message, there is a probability distribution over acts that is determined by the state of the receiver population, and hence a population-wide "receiver's strategy" instantiated for that message. The combination of these population-wide strategies plus the unconditional probability of states determine the mutual information between states and acts.

3 Results

Our main research question was how the presence of communication relates to common interest. We look into this question by generating a large number of random signaling games at each level of common interest, and then recording how likely it is in these games that random starting points evolve to a situation in which communication happens, depending on the level of common interest. Specifically, we focus on cheap talk signaling games with 3 equiprobable states, 3 messages and 3 acts. There are 10 possible values of C, our measure of common interest, for games of this sort. These games are individuated by 18 numbers: the 18 values in a payoff matrix of the form seen in Table 1.

For each value of C, we generated 1500 collections of 18 random integers between 0 and 99. Each one of these collections individuates a signalling game. A population of senders (the same applies to receivers) is characterized in terms of the frequencies of the 27 types of pure strategists who may be present. These pure strategies can be represented as follows, using the convention introduced in Table 2 in the introduction:

[4] cf. Zollman, Bergstrom and Huttegger (2013, 7). This can be done straightforwardly if, as in Table 2 above, strategies are rendered as matrices.

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

For each of the 1500 random games per value of C, we ran simulations starting from 1000 different randomly-chosen states of the two populations (the sender and receiver populations). This is equivalent to choosing 1000 random ordered pairs of points in the 26-dimensional simplex. At $t = 1000$, the pair of resulting population states was "translated" into a sender-receiver configuration, in the way described above, and the mutual information between states and acts was recorded. When the sender-receiver configuration at $t = 1000$ showed nonzero mutual information between states and acts, we scored that simulation as one in which communication evolves. In total, then, 10 (values of C) times 1500 (random games) times 1000 (pairs of random starting frequencies) simulations were run ($1.5 \cdot 10^7$ simulations).

We next note some implementation details. First, no effort was made to check whether by $t = 1000$ populations had settled into any specific kind of equilibrium or cyclic behavior, although casual inspection shows this to be often the case. Similarly, we have not checked for the stability, in any formal sense, of $t = 1000$ states; we have not assessed the consequences of small hypothetical deviations from these states.

Finally, to prevent rounding errors from impacting the results, we only count as nonzero amounts of information above 10^{-3} bits. It is possible, if unlikely, that rounding errors still play a role in the final results: calculations are carried out using 64-bit floating-point numbers. In the computer we have used, this means that the minimum frequency for a population is around $2 \cdot 10^{-308}$, types that go below this frequency simply becoming extinct. In the replicator dynamics it is impossible for a frequency to decline to exactly zero, so it is in principle possible that this computational limitation introduces a distortion: that is, it is in principle possible that types below the $2 \cdot 10^{-308}$ mark would have bounced back to nonnegligible frequencies. On the other hand, this number is so low that the empirical relevance of results that depended on types bouncing back from such a frequency would be doubtful.

A coarse-grained summary of the results of these simulations is given in figure 1. This figure shows, for each value of C, the overall proportion of simulations in which communication evolved. So we here group together, within each value of C, all the games with that value of C and all the initial states for each game. We then find, as shown in figure 1, that C is very predictive of this proportion: there are very few cases of the evolution of communication when $C = 0$ (although there *are* some; see Section 4 for discussion), while this is by far the most likely outcome when $C = 1$.[5] The dependence of communi-

[5]In our sample, there are 165 $C = 1$ games in which simulations *never* evolve to communication. These are all of the games in the sample (and the

cation evolution on C is monotonic and appears to be smooth across the chart (although bear in mind that there are only ten values of C, and curve fitting is therefore not entirely meaningful).

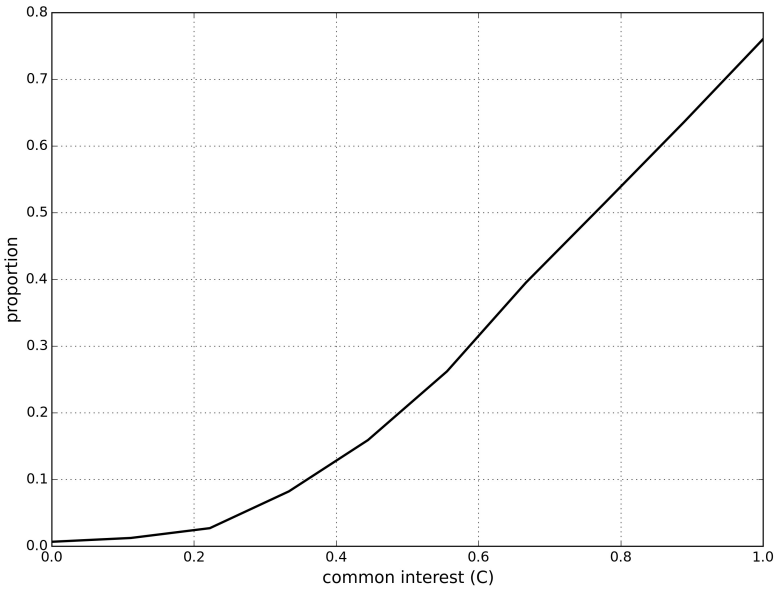

Figure 1: Proportion of simulations in which communication evolves, expressed as a function of common interest

only ones) in which one and the same act is the most preferred in every state. That is, in these games, if, e.g., act 3 is the most preferred for sender (and receiver: these are $C = 1$ games, so their preferences always coincide) in state 1, then it is also the preferred act in states 2 and 3. This makes communication useless: the receiver can ignore the sender's signals and simply do the best act no matter what. Godfrey-Smith and Martínez (2013) shows that the dependence of communication on this kind of *contingency of payoff* is as systematic as its dependence on the degree of common interest. However, Godfrey-Smith and Martínez (2013) errs in giving too strong a specification of the cases in which $C = 1$ games fail to accommodate communication; any $C = 1$ game in which the same act is best for every state will prevent communication, even if the value of other acts varies across states.

Apart from these systematic failures of communication, there are many $C = 1$ games in which a minority of simulations do not reach communication, but most other simulations for these games do.

In previous work, as noted in the introduction, we carried out an analysis of the prevalence of communicative Nash equilibria in samples of games with different values of C. This work used the same criterion for an "information-using" outcome that we employ here, though in the earlier paper this criterion only characterized equilibrium states. Our dynamic analysis in this paper confirms the conclusions drawn about the predictive value of C in the earlier work. In particular, the proportion of outcomes in which communication evolves, for each value of C, in the dynamic analysis is strongly correlated with the proportion of games, for each value of C, that contain an information-using Nash equilibrium as found in the earlier study. The Pearson correlation coefficient between the two series of values is 0.9990 (with $p < 0.001$).

Thus, the results of the dynamic model of this paper do appear to validate the findings of the earlier static analysis of the role of C in maintaining communication.

4 Communication at Very Low Values of C

As figure 1 shows, in some games with $C = 0$ we find a few starting points that evolve to situations in which communication is sustained by $t = 1000$. A value of $C = 0$ only obtains when the preferences of sender and receiver are reversed everywhere. That is, the most preferred act for the sender is the act least preferred by the receiver, and conversely, in every state. Remarkably, even in such a situation some simulations (ten thousand out of half a million runs, in our sample) can accommodate communication. On the other hand, this feature of our results does show some sensitivity to the dynamics chosen: when we ran the same simulations using the replicator-mutator dynamics, there were no runs in which a $C = 0$ game evolved to maintain communication at $t = 1000$. (Details of this analysis are given in Appendix A.)

A further notable feature of the dynamic results is that no simulation in our $C = 0$ sample evolved towards a Nash equilibrium in which information was being exchanged (although in some $C = 0$ games an information-using Nash equilibrium does exist.) This is evident from the fact that no such simulation was approaching an information-exchanging state (Nash or not) in which frequencies of behaviors had ceased to change by $t = 1000$. Instead, most of the $C = 0$ configurations at $t = 1000$ in which informative signaling is taking place belong to persisting cycles.[6] As an example, figure 2 shows the

[6]By "persisting cycle" we refer to a pattern in which the frequencies of types oscillate in an apparently stable manner. We have not, however, assessed the stability of these patterns beyond observations of dynamics up to $t = 1000$, and no conclusions should be drawn about nearby paths in the state space.

evolution of mutual information corresponding to a simulation based on the game in Table 3: very quickly, in a couple of hundred generations, the mutual information between states and acts enters a persisting cycle between 0.67 and 0.69 bits.

	S_1	S_2	S_3
A_1	31, 7	0, 95	57, 26
A_2	5, 71	99, 1	15, 62
A_3	17, 66	62, 23	28, 48

Table 3: A $C = 0$ game with persistently cyclical information exchange.

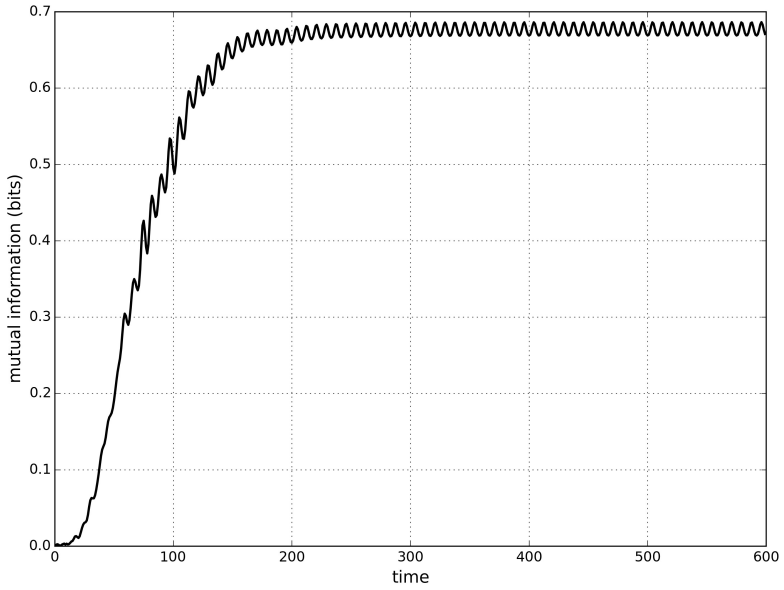

Figure 2: A $C = 0$ game evolving to cyclic communicative behavior.

Some other $C = 0$ games did not give rise to cycles, by $t = 1000$, but instead appeared to generate a chaotic dynamical regime. Results of this kind have also been found in a dynamic model of a Lewis signaling game by Wagner

(2012). Wagner used a stronger criterion for complete conflict of interest than $C = 0$ (he understood complete conflict of interest to exist only in constant sum games). Some of our simulations (such as the one corresponding to figure 3 and Table 4) often show communication at signaling-system levels. That is, the very incompatible preference rankings of sender and receiver can still sustain simulations in which the sender is, roughly half of the time, perfectly informative about the state of the world, and the receiver perfectly mindful of this information.[7]

	S_1	S_2	S_3
A_1	61, 28	14, 82	6, 74
A_2	11, 87	49, 58	7, 49
A_3	22, 71	21, 80	90, 38

Table 4: A $C = 0$ game with apparently chaotic orbits, in which information is often exchanged at signaling-system levels.

The signaling system that keeps recurring in the simulation corresponding to figure 3 is described in Table 5: the sender is perfectly informative, and the receiver exploits this to their benefit, and the sender's detriment – although the receiver does not carry the exploitation to the fullest extent; see below.

	M_1	M_2	M_3
S_1	1	0	0
S_2	0	1	0
S_3	0	0	1

	A_1	A_2	A_3
M_1	0	0	1
M_2	0	1	0
M_3	1	0	0

Table 5: The sender-receiver configuration in the signaling systems in figure 3.

What is sustaining informative signaling at $C = 0$ in these two kinds of cases (periodic orbits on the one hand, apparently chaotic orbits on the other)? There appears to be a main pattern in all of them: given that sender's and receiver's preferences are exactly reversed, the receiver would generally like to exploit any information in the messages sent by the sender – that is, use it to act in a way beneficial to them but detrimental to the sender. But any exploitation by the receiver will have to involve letting its behavior be guided

[7] In the simulation presented in figure 3, this behavior persists at least until $t = 60000$. We do not know wheter communication collapses at some later point.

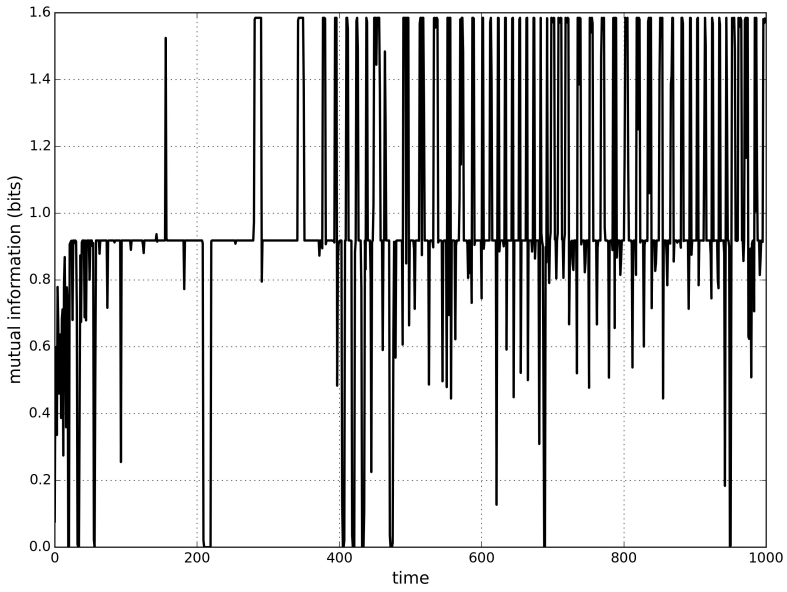

Figure 3: One evolution of mutual information between states and acts for the game presented in Table 4.

by the messages sent by the sender. This means the sender can exploit their attempted exploitation, by re-mapping states to signals in a way beneficial to them but detrimental to the receiver.

Godfrey-Smith (2013) describes one very simple kind of evolution that obeys this pattern: in a $C = 0$ game, if a sequential best-response regime is in place (the sender's strategy at t is the best response to the receiver's strategy at $t - 1$ which, in turn, is the best response to the sender's strategy at $t - 2$, etc.) and the sender kicks off the process by sending fully informative signaling (i. e., by using a strategy that is a bijection between S and M) then at any given time, the sender is mapping states to messages 1-to-1, and the receiver mapping messages to acts 1-to-1. Every configuration is a signaling system, but not as a result of good will: they are taking turns to exploit each other.

The persisting cycles in our $C = 0$ sample also appear to conform to this *sequential exploitation* pattern. Consider again the game in Table 3. Figure 4 is a fine-grained representation of the evolution of sender and receiver frequencies for the particular simulation that generated the results shown in figure 2; figure 2 shows change in mutual information while figure 4 shows change in the frequencies of the underlying behaviors.

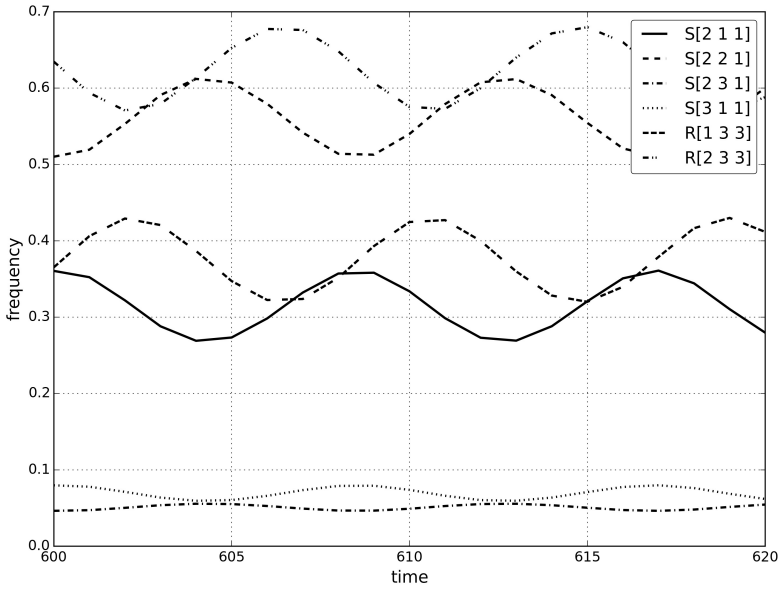

Figure 4: Evolution of frequencies in a persisting cycle for the game in Table 3

In what follows, we use S[M N O] as an abbreviation of the pure sender strategy consisting in sending (only) message M in S_1, message N in S_2, and message O in S_3. R[P Q R] stands for the pure receiver strategy that consists of doing (only) act P in response to M_1, act Q in response to M_2, and act R in response to M_3.

In the cycle, represented in figure 4, the only two types with nonzero frequencies in the receiver population are R[1 3 3] and R[2 3 3]. That is, the receiver always responds to M_2 and M_3 with A_3, and mixes A_1 and A_2 in response to M_1. The sender strategies with highest frequency are S[2 1 1] and S[2 2 1]. That is, the population mostly contains strategies that always send M_2 in S_1 and M_1 in S_3, but the strategies differ in how they respond to S_2, with the result that there is a population-wide mixing of M_1 and M_2 in response to that state. There is a small proportion of senders (below 10%) doing S[2 3 1] and S[3 1 1].

The frequencies of types R[1 3 3] and R[2 3 3] in the receiver's population change at the same rate as the proportions of S[2 1 1] and S[2 2 1] in the sender's population, only with a lag of approximately 3 time units. Here is what is going on: when the proportion of M_1 sent by the sender in S_2 falls, M_1 becomes more informative about S_3. In that case, the receiver wants to respond to M_1 with A_2, which secures the highest payoff for the receiver in S_3 (and is exploitative, insofar as the sender is then stuck with the lowest payoff in S_3). Consequently, the frequency of the R[2 3 3] type increases in the receiver population. As this frequency increases, the strategy consisting of sending M_1 in S_2 becomes more attractive for the sender (the pair S_2/A_2 has a very high payoff for them), and thus S[2 1 1] increases its frequency. Which again brings the frequency of R[1 3 3] up, etc.

Sender and receiver populations are thus launched in a cycle of sequential exploitation, but this is not all that is going on. For example, M_1 is, throughout, never sent in S_1. This gives a substrate of more cooperative and stable communication to the regime of attempted mutual exploitation: throughout the process, the receiver, when confronted with M_1 can rest assured that S_1 is not the case. Apparently, then, these cases should be understood in terms of a pair of phenomena, one cooperative and one non-cooperative. We outlined the role of exploitation above; here we will briefly attempt to characterize the second phenomenon, which involves a subtle form of cooperation. A sender and receiver can be seen as transforming, by means of mixed strategies, one game into another. Suppose that a sender sends M_1 always in S_1, mixes M_1 and M_2 in S_2, and never sends M_1 in S_3. Then when the sender sends M_1, they confront the receiver with an "uncertainty bundle"[8] that is partly com-

[8]We owe the "bundle" metaphor to Carl Bergstrom, and are grateful to both Bergstrom and Elliott Wagner for suggesting many of the outlines of the analysis given in these paragraphs. Rohit Parikh used a similar strategy in a treatment of the Crawford-Sobel model in an unpublished talk at the CUNY Graduate Center, October 2014.

prised of S_1 and partly comprised of S_2. The sender can be seen as giving the receiver perfect information about a bundle, rather than imperfect information about the "raw" state. A receiver, too, can create a bundle. When the receiver mixes A_1 and A_2, for example, in response to a given message, they present the sender with a bundle of acts that the sender must treat as a unit when they determine when to send that message.

Suppose, then, we revisit the game in Table 3 and consider the situation that obtains when the sender and receiver are following, for example, the strategies given in Table 6, which happen halfway through figure 4. We can redescribe this situation as one in which the sender is being perfectly informative about three "uncertainty bundles" they have constructed. The new "game" is shown in Table 7, where S^{M_i} stands for the bundle of states that the sender is presenting with message M_i. Here, sender and receiver agree on the worst act (the worst "raw" act, not the worst bundle of acts) possible in each state, and the new value of C is 0.66. Further, though, we can treat the receiver as creating bundles of acts: they are offering a new "act", A^{M_1}, which roughly consists of one third of A_1 and two thirds of A_2, and withdrawing access to the pure A_1 or A_2. If we re-interpret the game as transformed by both the sender's and receiver's bundling, we reach the payoff matrices shown in Table 8. This is now a "game" with complete common interest: both sender and receiver prefer A^{M_1} in S^{M_1}, and $A^{M_{2/3}}$ (the old A_3) in the other two state bundles. Because this "game" can be transformed again by either player changing their behaviors, and hence their bundling, the existence of common interest here does not have the same role that it has in an underlying game that acts as a fixed constraint. But we think that this description in terms of bundling may yield some understanding of how communication can arise in these apparently unlikely contexts.

	S_1	S_2	S_3
M_1	0	0.44	1
M_2	0.92	0.51	0
M_3	0.08	0.05	0

	M_1	M_2	M_3
A_1	0.36	0	0
A_2	0.64	0	0
A_3	0	1	1

Table 6: One sender-receiver configuration in the cycle represented in figure 4

It is more difficult to provide an intuitive description of what it is that forces senders and receivers into their behavior in the chaotic regimes that sometimes emerge in the game presented in table 4. A common pattern in the emergence of signaling systems in these regimes is quasi-periodic behavior in both sender and receiver, with two similar, but off-sync, "periods". For example, from $t = 600$ to $t = 620$ in the simulation corresponding to figure 3, the sender alternates S[2 1 1] with S[1 2 3]. Meanwhile, the receiver alternates R[3

	S^{M_1}	S^{M_2}	S^{M_3}
A_1	39.58, 47.08	19.94, 38.38	19.08, 40.85
A_2	40.67, 43.36	38.52, 46.03	41.15, 44.08
A_3	38.39, 40.36	33.04, 50.66	34.31, 49.46

Table 7: The game in Table 3, as rebundled by the sender.

	S^{M_1}	S^{M_2}	S^{M_3}
A^{M_1}	40.28, 44.70	31.84, 43.28	33.21, 42.91
$A^{M_{2/3}}$	38.39, 40.36	33.05, 50.66	34.31, 49.46

Table 8: The game in Table 3, as rebundled by sender and receiver.

2 1] with R[1 2 2]. As the sender's "cycle" is out of sync with the receiver's, all four combinations occur: S[2 1 1] and R[3 2 1]; S[1 2 3] and R[3 2 1]; S[2 1 1] and R[1 2 2]; S[3 2 1] and R[1 2 2]. The second among these combinations is a signaling system. The other three are partially informative. Thus the 1.58 bits/0.91 bits alternation in figure 3 between $t = 600$ and $t = 620$.

In fact, by the time senders and receivers engage in this behavior, the frequencies of most sender and receiver types is zero; in particular, the reason why the receiver doesn't respond to S[1 2 3] with R[2 1 1] is that this type is extinct by $t = 600$. The best the receiver can do is engage in signaling-system behavior, with R[3 2 1]. What we see is sequential exploitation to the full extent of their current capabilities. In the case we have been discussing, the difference between frequencies that are very close to zero and those that are effectively zero (below $2 \cdot 10^{-308}$; see above) turns out to be very important: as figure 5 shows, frequencies jump to extremely close to zero to extremely close to one, and, if we round population frequencies so that every type with frequency below 10^{-10} is declared extinct, signaling systems fail to appear.

The above analysis, in any case, only focuses on one particularly legible fragment of the simulation. Most of what happens in the full run depends on haphazard details of the population structure, as is bound to happen in apparently chaotic behavior of this sort, and could not be convincingly labelled as sequential explotation. The unpredictable alternation of quasi-periodic patterns that can be noticed in a longer run is shown in figure 6, which presents the evolution of the sender-receiver configurations to which the population structures translates between $t = 600$ and $t = 800$.

As we have said, none of the $C = 0$ runs in our sample in which communication is maintained evolves towards a Nash equilibrium. In fact, in our sample, the first case where there is evolution towards a Nash equilibrium happens at $C = 0.22$. Somewhat surprisingly, the situation appears to be at follows. On one side, we have anecdotal evidence to the effect that when a $C = 0$ game does have a Nash equilibrium in which communication is maintained, then it is very likely that some initial population frequencies will lead to persisting

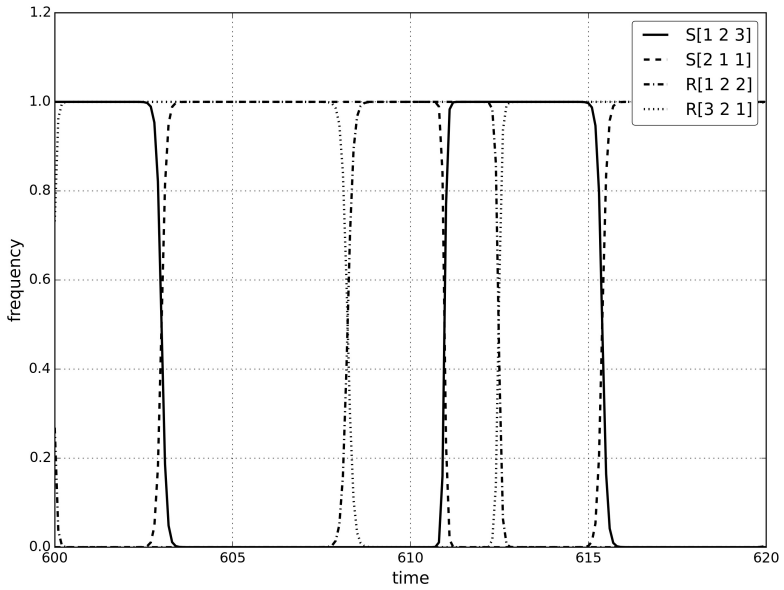

Figure 5: Detail of the evolution of populations corresponding to figure 3, between $t = 600$ and $t = 620$.

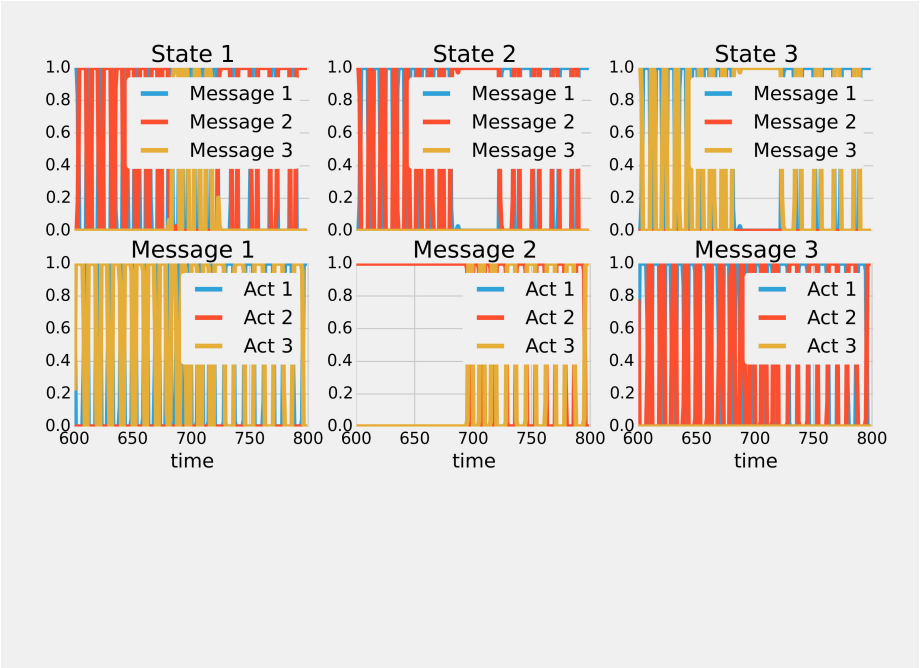

Figure 6: Evolution of sender-receiver configurations corresponding to figure 3, between $t = 600$ and $t = 800$.

communication: in all of the 24 such games we have found in the random sampling prepared for this and our previous paper, the replicator dynamics will take some initial conditions to a state at $t = 1000$ in which communication does then persist. On the other side, in no case that we have found there is evolution towards the Nash equilibrium itself.

5 Conclusions

This paper describes the results of a dynamic analysis of the role of common interest in the evolution of communication in a three-state Lewis signaling game. We find a strong predictive role for common interest, as measured by C: in a large random sample of games, the proportion of evolutionary simulations in which communication was maintained at $t = 1000$ was monotonically associated with C. The results presented here complement those in an earlier paper (Godfrey-Smith and Martínez 2013) which gave a purely static analysis of games of this kind, also using C as a measure of common interest. The two sets of results are broadly consistent and complementary; each approach provides a different perspective on these systems.

First, an analysis using Nash equilibria can be used to give a coarse-grained description of a range of systems which operate under different dynamical regimes, and also cases involving human choice where no well-defined "dynamic" may exist at all. The replicator-dynamic model, on the other hand, gives a much finer-grained representation of systems to which it applies, and has been shown to be informative about systems that follow a somewhat different dynamic, as well (Bendor and Swistak 1998).

Our work here also uncovers phenomena that involve cycling and apparently chaotic behaviors. We offered an initial analysis of these outcomes in terms of a combination of sequential exploitation and the transformation of games through "uncertainty bundling," but this last analysis was offered briefly, as a first foray; clearly much more work remains to be done on the interaction between common interest and evolutionary dynamics in signaling games.

References

Bendor, Jonathan, and Piotr Swistak. 1998. "Evolutionary equilibria: Characterization theorems and their implications.", *Theory and Decision*, 45 (2):99-159.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic information transmission", *Econometrica*, 50:1431-1451.

Godfrey-Smith, Peter. 2013. "Information and Influence in Sender-Receiver Models, With Applications to Animal Communication", in *Animal commu-*

nication theory: Information and influence, ed. Ulrich Stegmann, 377-396. Cambridge: Cambridge University Press,

Godfrey-Smith, Peter, and Manolo Martínez. 2013. "Communication and Common Interest", *PLOS Computational Biology*, 9 (11): e1003282.

Grafen, Alan. 1990a. "Sexual Selection Unhandicapped by the Fisher Process", *Journal of Theoretical Biology*, 144: 473-516.

Grafen, Alan. 1990b. "Biological Signals as Handicaps", *Journal of Theoretical Biology*, 144: 517-546.

Grice, Paul. 1957. "Meaning", *Philosophical Review*, 66: 377-388.

Hofbauer, Josef, and Karl Sigmund. 1998. *Evolutionary games and population dynamics*, Cambridge: Cambridge University Press.

Huttegger, Simon M. 2007. "Evolutionary explanations of indicatives and imperatives", *Erkenntnis*, 66 (3): 409-436.

Lachmann, Michael, Szabolcs Szamado, and Carl T. Bergstrom. 2001. "Cost and conflict in animal signals and human language". *Proceedings of the National Academy of Sciences*, 98 (23): 13189-13194.

Lewis, David. 1969/2002. *Convention*, Oxford: Wiley-Blackwell.

Martínez, Manolo. 2015. "Deception in sender-receiver games", *Erkenntnis*, 80, (1): 215–227.

Maynard-Smith, John, and David Harper. 2003. *Animal signals*, Oxford: Oxford University Press.

Millikan, Ruth G. 1984. *Language, Thought and Other Biological Categories*, Cambridge: The MIT Press.

O'Connor, Cailin. 2014. "The evolution of vagueness", *Erkenntnis*, 79 (4): 707-727.

Sandholm, William H. 2010. *Population games and evolutionary dynamics*, Cambridge: The MIT Press.

Shannon, Claude E. and Warren Weaver. 1949. *The Mathematical Theory of Communication*, Champaign: University of Illinois Press

Skyrms, Brian. 1996. *Evolution of the Social Contract*, Cambridge: Cambridge University Press.

Skyrms, Brian. 2010. *Signals: Evolution, Learning & Information*, New York: Oxford University Press.

Spence, Michael. 1973. 'Job Market Signaling', *Quarterly Journal of Economics*, 87 (3): 355-374.

Sterelny, Kim. 2012. *The Evolved Apprentice: How Evolution Made Us Unique*, Cambridge: The MIT Press.

Tomasello, Michael. 2008. *Origins of human communication*, Cambridge The MIT Press.

Wagner, Elliott O. 2014. "Conventional semantic meaning in signalling games with conflicting interests", *British Journal for the Philosophy of Science*. doi: 10.1093/bjps/axu006.

Wagner, Elliott O. 2012. "Deterministic chaos and the evolution of meaning", *British Journal for the Philosophy of Science*, 63: 547-575.

Zahavi, Amotz. 1975. "Mate Selection: A Selection for a Handicap", *Journal of Theoretical Biology*, 53: 205-214.

Zollman, Kevin. 2011. "Separating directives and assertions using simple signaling games", *The Journal of Philosophy*, 108 (3): 158-169.

Zollman, Kevin, Carl T. Bergstrom, and Simon M. Huttegger. 2013. "Between cheap and costly signals: The evolution of partial honest communication", *Proceedings of the Royal Society B*, 20121878.

Appendices

A Replicator-Mutator Dynamics

In the main text, communication evolution has been shown to depend on C in simulations in which evolution is governed by the two-population replicator dynamics in continuous time. A very similar situation is observed if, instead, we let populations evolve according to the two-population *replicator-mutator* dynamics. In this alternative, the rate of change of the frequency of a certain type depends not just on how well it does compared to the average in its population, but also on a *mutation matrix*, M, each member M_{ij} of which gives the probability that an individual of type i changes its type to j. The differential equations for the replicator-mutator dynamics are, thus, as follows:

$$\dot{x}_i = \sum_j (x_j M_{ji} \overline{\pi}_i^{\sigma}) - x_i \overline{\pi}^{\sigma} \tag{A1}$$

$$\dot{y}_i = \sum_j (y_j M_{ji} \overline{\pi}_i^{\varrho}) - y_i \overline{\pi}^{\varrho} \tag{A2}$$

In our simulations we have used a mutation matrix according to which types "breed true" with high probability, and mutate equiprobably to every other type. That is, for a population of 27 pure-strategist types (such as the sender and receiver populations in our model), M is of the following form (with $m = 0.005$):

$$M = \begin{pmatrix} 1-m & \frac{m}{26} & \cdots & \frac{m}{26} \\ \frac{m}{26} & 1-m & \cdots & \frac{m}{26} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{m}{26} & \frac{m}{26} & \cdots & 1-m \end{pmatrix}$$

Again here, the proportion of simulations that show evolution to communication increase monotonically and smoothly with C. See figure A1.

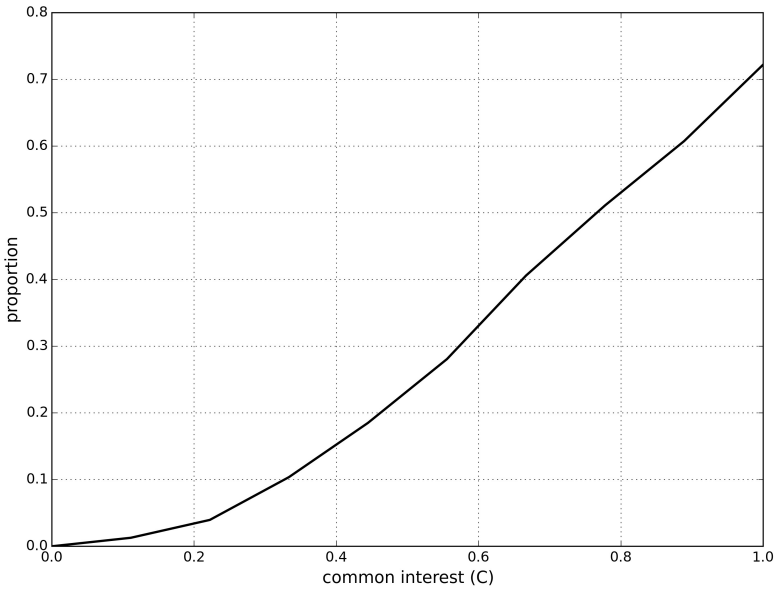

Figure A1: Proportion of simulations in which communication evolves per value of common interest. Populations evolve according to the replicator-mutator dynamics

One important difference between the replicator and replicator-mutator results is that, in the latter, no simulation in the $C = 0$ group presents evolution to communication. It should be noted, though, that M is such that everything mutates to everything else; the net effect of this mutation is the introduction of a certain amount of "noise" which, among other things, prevents type frequencies from dropping below $\frac{m}{26}$. Whether communication at $C = 0$ would be possible in the presence of a more structured mutation matrix remains to be seen.

B Other Implementation Details

We have implemented our simulations in custom scripts relying on the Python scientific stack: Python 3.4.2, NumPy 1.9.1, and SciPy 0.15.1. The systems of ordinary differential equations were solved using the scipy.integrate.odeint solver, which, in its turn, calls the LSODA solver of the ODEPACK library (see http://www.netlib.org/odepack/opkd-sum for details). Whenever this solver failed, our scripts fell back to the implementation of the Dormand-Prince method (Dormand and Prince 1980) provided by scipy.integrate.ode. Figures were prepared with matplotlib 1.4.3. Our scripts are published under the GPL license at https://github.com/manolomartinez/signal.

The random sampling of population starting points followed the 27-dimensional flat Dirichlet distribution, calculated using the NumPy implementation available in numpy.random.dirichlet().

Additional Reference

Dormand, John R., and P. J. Prince. 1980. 'A family of embedded Runge-Kutta formulae', *Journal of Computational and Applied Mathematics*, 6, (1): 19-26.