bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

THE EVOLUTION OF COVERT SIGNALING

PAUL E. SMALDINO[1], THOMAS J. FLAMSON[1], AND RICHARD MCELREATH[2]

April 28, 2017

Human sociality depends upon the benefits of mutual aid and extensive communication. However, mutual aid is made difficult by the problems of coordinating diverse norms and preferences, and communication is harried by substantial ambiguity in meaning. Here we demonstrate that these two facts can work together to allow cooperation to develop, by the strategic use of deliberately ambiguous signals: covert signaling. Covert signaling is the transmission of information that is accurately received by its intended audience but obscured when perceived by others. Such signals may allow coordination and enhanced cooperation while also avoiding the alienation or hostile reactions of individuals with different preferences. Although the empirical literature has identified potential mechanisms of covert signaling, such as encryption in humor, there is to date no formal theory of its dynamics. We introduce a novel mathematical model to assess the conditions under which a covert signaling strategy will evolve, as well as how signaling strategies coevolve with receiver attitudes. We show that covert signaling plausibly serves an important function in facilitating within-group cooperative assortment by allowing individuals to pair up with similar group members when possible and to get along with dissimilar ones when necessary. This mechanism has broad implications for theories of signaling and cooperation, humor, social identity, political psychology, and the evolution of human cultural complexity.

Keywords: cooperation, signaling, social identity, ethnic markers, humor, homophily

---

[1] Cognitive and Information Sciences, University of California, Merced, Merced, CA, USA

[2] Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

E-mail address: `paul.smaldino@gmail.com`.

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

INTRODUCTION

Much of the research on human cooperation has focused on the free-rider problem: how to maintain cooperation when individuals' immediate interests are opposed to those actions that would maximize the total benefits to the group. However, in many cases individuals' interests are *aligned* rather than opposed, and these mutualistic scenarios may be equally important in understanding human social evolution (Skyrms, 2004; Calcott, 2008; Tomasello et al., 2012; Smaldino, 2014). Even though there are no incentives for individuals to defect, mutualism is still a dilemma. When individuals *differ* in preferences, norms, or goals, the ability to efficiently coordinate breaks down. Therefore coordinating behavior, and in particular forming the reliable expectations of partner behavior that make coordination possible, is essential for the evolution of mutualism (Schelling, 1960).

Take, for example, the Battle of the Sexes game (Luce & Raiffa, 1957), in which there exist two equivalent Nash equilibria. Each player still prefers to coordinate rather than go it alone, but has a different idea of how to coordinate best. Each player would be better off finding another co-player with more aligned interests. Human societies are replete with dilemmas of this kind (Boyd & Richerson, 1994), and the need to coordinate extends to other forms of collective action as well (Ostrom, 2000). Institutional mechanisms like punishment effectively convert other social dilemmas into coordination dilemmas, expanding their importance for understanding human sociality.

How can individuals assort on the basis of similarity in preferences, norms, goals, or strategies? Often these traits are difficult or impossible to directly observe. When preferences and norms are consciously held, individuals can merely signal their preferences. But often individuals are not conscious of their norms and preferences or realize their relevance too late to signal them.

One solution is the evolution of ethnic marking. Anthropologists have long recognized the importance of ethnic markers or tags in signaling group membership to improve cooperative outcomes (Barth, 1969), and in recent years an extensive formal literature has developed exploring how these arbitrary signals can facilitate assortment on unconscious norms and preferences (Nettle & Dunbar, 1997; Castro & Toro, 2007; Efferson et al., 2008; Mace & Holden, 2005; McElreath et al., 2003; Moffett, 2013).

Language style and content can serve as a marker for social coordination (Nettle & Dunbar, 1997). However it is striking that much communication is ambiguous. Is this ambiguity merely the result of constraints on the accuracy of communication? Here we propose that such ambiguity may serve to facilitate coordination and thereby enhance cooperation within human societies. A basic problem with unambiguous signals is that they may foreclose less coordinated partnerships that may be of value in different contexts. In some situations, such as frequent or long-term endeavors, one is best served by engaging in homophilic assortment with a small set of similar partners who afford relatively more efficient coordination of behavior. In other contexts, different assortment outcomes may be desired, such as larger-scale cooperation for communal defense, differently-skilled partners for gains in trade, or when vying for the assistance of high-status individuals in political advancement. While overt signals of personal qualities like ethnic markers are useful in some contexts, where the adaptive problem is to delimit a set of partners who subscribe to

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the same broad behavioral norms and to categorically avoid interaction with those who do not, the "all-or-nothing" character of such signals makes them inadequate for dealing with assortment *within* the groups delineated by these markers. Individuals will often benefit from not "burning bridges" with less similar group members, so as to be able to draw on them for cooperation in other contexts. A signaling system that enables group members to communicate relative similarity only when similarity is high while retaining a shroud of ambiguity when similarity is low would facilitate assortment when the situation allows for it and still enable low-similarity assortment when the situation demands it.

Here we analyze the evolution of cooperative assortment by a form of signaling which satisfies these requirements and is known empirically to exist in nearly all human societies. Covert signaling is the transmission of information that is accurately received by its intended audience but obscured when perceived by others. It may naïvely appear that communication should have clarity as its goal. However, purposeful ambiguity is often strategic, allowing signalers both flexibility and plausible deniability (Eisenberg, 1984; Pinker et al., 2008). Leaders may use ambiguous language to rally diverse followers under a common banner (Eisenberg, 1984), politicians may use vague platforms to avoid committing to specific policies (Aragonès & Neeman, 2000), and would-be suitors may mask their flirtations to be viewed innocuously if their affections are unreciprocated (Gersick & Kurzban, 2014). What these discussions of ambiguity have in common is the assumption that all receivers will find the signals to be equally vague. In contrast, our discussion focuses on signals that will be clearer for some receivers and more ambiguous for others. A common example is "dog-whistle politics" (López, 2014), in which statements have one meaning for the public at large and a more specialized meaning for others. Such language attempts to transmit a coded message while alienating the fewest listeners possible.

A more precise and possibly much more common form of covert, within-group signaling is humor (Flamson & Barrett, 2008; Flamson & Bryant, 2013). According to the encryption model of humor, a necessary component of humorous production is the presence of multiple, divergent understandings of speaker meaning, some of which are dependent on access to implicit information. Only those listeners who share access to this information can "decrypt" the implicit understandings and understand the joke. Because the successful production of a joke requires access to that implicit information, humor behaves in manner similar to "digital signatures" in computer cryptography, verifying the speaker's access to that information without explicitly stating it. By not explicitly declaring one's position within local variation, but instead signaling and assessing similarity on the basis of subtle and iterated cues that only like-minded group members can detect, individuals can engage in positive assortment in some contexts without burning bridges with potential allies in others. While not all humor necessarily has this form or function, a substantial amount of spontaneous, natural humor does (Flamson & Barrett, 2008; Flamson & Bryant, 2013).

Covert signaling can also facilitate assortment along dimensions of similarity more nuanced than discrete types or groups, which is a common restriction of ethnic markers or tags as they are typically discussed (McElreath et al., 2003; Antal et al., 2009; Cohen & Haun, 2013; Hammond & Axelrod, 2006). Jokes and other encrypted signals of identity can convey rich information about an individual's beliefs, goals,

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

personality, proclivities, and history. Although any two individuals within a group should be able to cooperate when it is mutually beneficial to do so, pairs who are more similar along these trait dimensions should cooperate more effectively, generating larger benefits.

We propose that, by avoiding burned bridges, covert signaling serves an important function in facilitating cooperative assortment within groups by allowing individuals to pair up with similar individuals when possible and with dissimilar individuals when necessary. In the remainder of this paper, we precisely define the strategic logic of covert signaling in the context of opportunities for social assortment and coordination. We analyze the conditions for covert signaling to be preferred over overt signaling, in which information about an individual's traits is more transparent. Covert signaling is not always favored. For example, if it is possible to freely choose cooperative partners from a very large pool, overt signaling may be a more advantageous means of communication, as individuals will both avoid being paired with dissimilar partners and reap the added benefits that comes from *knowing* that a similarity exists (Chwe, 2001). However, covert signaling often will be favored. It sacrifices maximal transparency for the sake of maintaining working relationships with dissimilar individuals. Although covert signals will often be noisier than overt signals, and therefore will be received with reduced accuracy, we will show that such increased noise—and therefore increased ambiguity—can in some cases be advantageous. Covert signaling therefore may be an important part of a full explanation of both specific forms of communication and coordination, such as coded speech and humor, as well as the flexibility of human sociality more generally. Our model also yields specific predictions about default attitudes toward strangers in the absence of clear signals, with implications for understanding differences between contemporary political affiliations.

Model Description

We consider a large population of individuals who are generally cooperative toward one another. That is, we assume they have already solved the first-order cooperation problem of suppressing free riders, and can instead focus on maximizing the benefit generated by cooperation (Skyrms, 2004; Calcott, 2008; Tomasello et al., 2012; Smaldino, 2014). Although individuals all belong to the same large group, they vary along many trait dimensions and thus share more in common with some individuals than others. Pairs of individuals whose trait profiles overlap to some threshold degree are deemed *similar* (S). Otherwise they are deemed *dissimilar* (–S). Pairs of similar individuals can more effectively coordinate, and so can obtain higher payoffs from cooperation. The probability that two randomly selected individuals will have similar trait profiles is given by s.

Our model proceeds in discrete generations, and each generation is subdivided into two stages. In the first stage, individuals signal information about their trait profiles to the other members of the group. In the second stage, individuals interact in one of two ways and receive payoff conditional upon attitudes formed in the first stage.

Stage 1: Signaling. In the first stage, individuals may produce either an overt or covert signal of their underlying traits. Overt signals are received by a fraction R of the population and explicitly signal similarity or dissimilarity. Covert signals

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

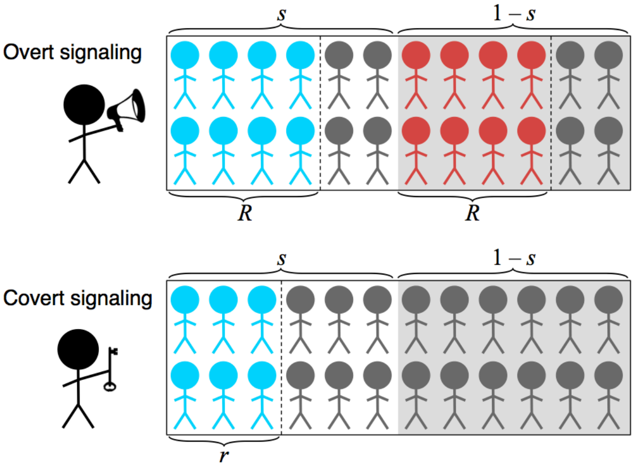

FIGURE 1. Illustration of signaling dynamics. For any signaler, a proportion s of the population is similar. For overt signalers, a proportion R will receive the signal. Of those, similar individuals will like them (blue) and dissimilar individuals will dislike them (red). Anyone not receiving the signal will remain neutral (gray; but see also our analysis where attitudes default to dislike in the absence of signals). For covert signalers, a smaller proportion of similar individuals will receive the signal (r), but dissimilar individuals will not.

in contrast are received by a fraction $r < R$ of the population and have content contingent upon similarity of the sender and receiver. See Figure 1. When the sender and receiver are similar, covert signals are received as signaling *similar*. Otherwise, the receiver does not notice the signal and acts as if a signal were not received at all. We allow a family of continuous signaling strategies in which covert signals are produced a fraction p of the time.

Receivers have both a default attitude towards all individuals in the population and update this attitude conditional upon received signals. This implies that receiver strategies must map three signal states—similar, dissimilar, no signal—to an attitude. We consider three discrete attitudes: like, dislike, and neutral. These three discrete attitudes correspond to the hypothesis that covert signals help individuals to avoid being disliked while also achieving sufficient positive assortment by type. Two attitudes would be too few, because it would force agents to adopt either like or dislike as a default attitude, removing any incentive for covert signals. Three is the minimum required to model the hypothesis.

We allow a continuous family of receiver strategies that probabilistically map signals to attitudes. Each strategy parameter a_{XY} indicates the probability of mapping signal X $\in$ {Similar,None,Dissimilar} to attitude Y $\in$ {Like,Neutral,Dislike}. The total receiver strategy can be represented by a table:

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

	Like	Neutral	Dislike
Similar	a_{SL}	a_{SN}	a_{SD}
None	a_{NL}	a_{NN}	a_{ND}
Dissimilar	a_{DL}	a_{DN}	a_{DD}

The three parameters in each row are constrained to sum to one.

Stage 2: Interaction. After attitudes are established, pairs of individuals interact. There are two interaction contexts, each with its own mode of dyad formation: *free choice* and *forced choice* scenarios. In a *free choice* scenario, dyads form conditional on the attitudes of both individuals. In contrast, in a *forced choice* scenario, an individual must seek help from whomever happens to be around and dyads are not conditional upon shared attitudes. Under these circumstances, it may be important not to have burned bridges, since this will limit the likelihood of effective coordination. These contexts are starkly different. Real contexts are often some compromise between these extremes. We use these contexts to present the clearest investigation of the hypothesis that covert signals trade off worse performance in assortment contexts, like the *free choice* context, for the ability to avoid being disliked in *forced choice* contexts.

In the free choice context, dyads form in proportion to the joint attitudes of each pair of individuals, but attitudes do not directly influence payoffs. Instead, underlying similarity influences payoffs. Specifically, similar dyads receive an average payoff of 1, establishing a baseline measurement scale. Dissimilar dyads receive a payoff of zero. We assume that each individual in a dyad who *likes* the other individual increases the proportional odds of that dyad forming by a factor $w_{\text{L}} > 1$. For each individual who *dislikes* the other, the proportional odds of the dyad forming are reduced by a factor $w_{\text{D}} < 1$. This implies five possible kinds of dyads that might interact: LL, LN, NN, ND, and DD. We assume that the proportional odds of each, relative to random assortment, are: w_{L}^2, w_{L}, 1, w_{D}, and w_{D}^2. These parameters are fixed features of the social environment, not aspects of strategy. This prevents strategy dynamics from generating perfect assortment. Receiver strategies that assign attitudes in ways that make good use of signal information will achieve better assortment and receive higher payoffs, conditional on the assortment constraints determined by w_{L} and w_{D}.

In the forced choice context, dyads form at random with respect to attitudes, but attitudes do instead directly influence payoffs. This context entails a baseline payoff of 1 for both individuals. However, attitudes adjust payoffs, because negative attitudes make it harder to interact. When one individual dislikes the other, he makes the interaction more difficult than it must be and thereby imposes a cost $-d$ on the other individual. When both individuals dislike one another, their difficulties act synergistically, inducing an additional cost $-\delta$ on each. As we show later, these synergistic costs are very important to the overall signaling dynamics.

Let q be the relative importance of the free choice context and $1-q$ the relative importance of the forced choice context.

Payoff expression. With the assumptions above, we can define a general payoff expression for a rare individual with signal strategy p' and receiver strategy matrix $\mathbf{a}'$ in a population with common-type strategy $\{p, \mathbf{a}\}$. The expected payoff to this

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

individual is:

$$W(p', \mathbf{a}') = \Omega + q \Pr(\text{similar}|p', \mathbf{a}') \\ + (1-q)\big(1 - \Pr(\text{disliked}|p', \mathbf{a}')(d + \Pr(\text{dislike}|p', \mathbf{a}')\delta)\big) \qquad (1)$$

where Ω is an expected baseline payoff due to other activities. The work lies in defining the probabilities $\Pr(\text{similar}|p', \mathbf{a}')$, $\Pr(\text{dislike}|p', \mathbf{a}')$, and $\Pr(\text{disliked}|p', \mathbf{a}')$. In the mathematical appendix, we show how to define these probabilities, using the assumptions above. The resulting general payoff expression is very complicated. In the following section, however, we are nevertheless able to analyze it by considering invasion and stability of relevant combinations of signaling strategy and receiver strategy.

Analysis and Results

The motivating hypothesis is that covert signals can proliferate because they allow sufficient assortment in the free choice context and also reduce being disliked in the forced choice context. To evaluate the logic of this idea, we proceed by asking when covert signals can be stable, when they can invade, and which receiver strategies are necessary for their stability or invasion. We are able to demonstrate that the following conditions favor covert signals.

(1) Covert signals require a sufficient proportion of receivers to default to neutral attitudes. If everyone defaults to disliking, then covert signals can produce no benefit. Defaulting to neutral is favored under a wide range of conditions, provided that covert signals are sufficiently hard to receive (r is not too large) and avoidance of disliked individuals is not too efficient (w_{D} is not too small).

(2) Covert signals require that the cost of being disliked in the forced choice context be sufficiently high. This also means that baseline similarity in the population (s) must be sufficiently low, because this creates the risk of being disliked by dissimilar individuals.

(3) Overt signalers cannot have too large an advantage in the free choice context. This requires that assortment with liked individuals not be too accurate. The accuracy of assortment is influenced by the reception probabilities of both signal types, R and r, as well as the proportional odds assortment factors, w_{L} and w_{D}.

In the remainder of this section, we derive these results and provide intuition for why they hold. First we derive simple evolutionary dynamics for these payoffs. This allows us to submit the model to invasion and stability analysis, asking both when covert signals can be stable and when they may invade a population of overt signals. Then we proceed by considering the dynamics within each interaction context—the forced choice context and the free choice context—separately. The model is much easier to understand this way, as each context induces unique incentives for signalers and receivers. Then we summarize the joint dynamics of the full model with both contexts.

Evolutionary dynamics. We generate evolutionary dynamics for the strategy space by assuming that rare invading strategies increase in frequency when they achieve higher payoffs than a common-type strategy. A number of different biological assumptions can generate such dynamics. For example, naive individuals each

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

generation could learn their strategies from successful individuals in the previous generation. Or an individual in overlapping generations could update strategies each generation by comparing its own payoff to that of another, random member of the population. Genetically coded strategies that influence biological fitness would also generate this dynamic. We remain agnostic about inheritance and transmission mechanism, because the point of our modeling exercise is to explore the design aspects of covert signals. This is best achieved by a form of analysis that abstracts away from transmission details, even though of course in any real system such details will turn out to influence which strategies are possible and how they evolve.

The evolutionary dynamics are determined by evolution in both p and the attitude parameters, so we define selection gradients for each of these parameters. The gradient for signaling is defined by:

$$g(p) = \left.\frac{\partial W(p', \mathbf{a}')}{\partial p'}\right|_{p'=p, \mathbf{a}'=\mathbf{a}} \tag{2}$$

The gradients for each receiver parameter is defined similarly.

The potential space of receiver strategies is very large. However, the relevant space of strategies is fairly small. In the mathematical appendix, we show that payoff dynamics always favor mapping *similar* signals to *like* attitudes, implying $a_{\text{SL}} = 1$. The reason is that maximizing probability of assortment for similar individuals maximizes payoffs, and the *like* attitude maximizes the probability of interacting in the free choice context. On the other hand, payoff dynamics do not always favor mapping *dissimilar* signals to *dislike* attitudes. The reason is that the forced choice context disfavors disliking whenever $\delta > 0$. We therefore constrain further analysis to the relevant situations in which the penalty for mutual dislike, δ, is small enough that assortment incentives favor mapping *dissimilar* signals to *dislike* attitudes. We reemphasize this important constraint on the relevance of covert signals in the discussion, because constraints of this sort help in producing predictions. Finally, with respect to default attitudes, formed when no signal is received, payoff dynamics never favor assigning *like*, because this erodes the value of assigning *like* to *similar* signals.

The remaining default receiver parameters are free to evolve. Therefore, for most of the analysis to follow, we assume that $a_{\text{SL}} = 1$, $a_{\text{DD}} = 1$, $a_{\text{NL}} = 0$, $a_{\text{NN}} = 1 - \alpha$, and $a_{\text{ND}} = \alpha$. This allows us to use the gradient on α, defined by:

$$g(\alpha) = \left.\frac{\partial W(p', \alpha')}{\partial \alpha'}\right|_{p'=p, \alpha'=\alpha} \tag{3}$$

to ask when evolution favors assigning *dislike* to no signal, $\alpha > 0$, effectively defaulting to disliking everyone. It will be convenient to refer to $\alpha = 0$ as the *generous receiver* strategy and $\alpha = 1$ as the *churlish receiver* strategy.

Dynamics of the forced choice context. We begin by analyzing the incentives and evolutionary dynamics of the forced choice context. In this context, incentives favor covert signals, because such signals are better at avoiding being disliked. However, this advantage depends upon incentives favoring generous receiver strategies that do not dislike by default. Luckily for covert signals, receiver incentives in this context always favor generous receiver strategies as long as there is any negative synergy, $\delta > 0$. Therefore, we show below, the forced choice context favors generous receivers, $\alpha = 0$, which in turn favor covert signals, $p = 1$.

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The gradients in this context are:

$$g(p)|_{q=0} = \big(\alpha r s + R(1-\alpha-s)\big)\big(d + \delta(\alpha(1-prs) + R(1-p)(1-\alpha-s))\big) \quad (4)$$

$$g(\alpha)|_{q=0} = -\delta\big(1-(1-p)R-prs\big)\big(\alpha(1-prs)+(1-p)R(1-\alpha-s)\big) \quad (5)$$

These expressions seem complex at first, but produce fairly simple dynamics. First let's ask when p can increase. When $\alpha = 0$, the generous receiver strategy is common, and covert signals can increase when:

$$d + \delta(1-s)(1-p)R > 0 \quad (6)$$

This is satisfied for any allowable values of the parameters. Note also that it does not require both a direct cost of being disliked, d, and a synergistic cost of mutual dislike, δ. Either one is sufficient to favor covert signals, as long as α is small. Next consider when $\alpha = 1$, the churlish receiver strategy is common. Then covert signals can increase when:

$$d + \delta\big(1 - s(pr + (1-p)R)\big) < 0 \quad (7)$$

And this is never satisfied, for any p. Therefore covert signals are favored when $1-\alpha$, the amount of generous receiver behavior, is sufficiently high. The threshold value is found where $g(p) = 0$:

$$\hat{\alpha} = \frac{1-s}{1-\frac{r}{R}s} \quad (8)$$

When α is above this value, overt signals are favored. When it is below it, covert signals are favored. Why? When receivers are relatively generous, and there is sufficient dissimilarity in the population, covert signals reduce costs by avoiding being disliked. If generous receivers are relatively rare, however, then covert signals can actually do worse than overt signals, because they are received less often than overt signals, $r < R$. If r/R is sufficiently small, overt signals are favored for a wide range of values of α. If however $r = R$ and covert signals have no disadvantage in audience size, then overt signals are never favored in this context, no matter the amount of similarity s.

Now the crucial question is when α will fall below this threshold $\hat{\alpha}$. The condition for payoff dynamics to favor smaller values of α, in the forced choice context only, is:

$$\delta > 0 \quad (9)$$

Therefore the forced choice context always favors smaller values of α, provided there is any negative synergy between disliking and being disliked. Otherwise α is neutral and does not move at all, based on payoff dynamics. Why does this context always favor generous receivers? There is no advantage to be had in disliking people in this context, because assortment does not depend upon attitudes. Payoffs depend upon attitudes, however, and mutual dislike results in poor payoffs. Therefore, it pays to be generous in attitudes towards those one has no information about.

While this context favors covert signals, whether or not the full system favors such signals will depend upon the strength with which it favors them, as well as the incentives in the free choice context and the importance of each context to total payoffs.

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Dynamics of the free choice context. In the free choice context, attitudes influence assortment but do not directly influence payoffs. Instead, hidden norm similarity influences payoffs. The free choice context favors overt signals over covert signals, because overt signals increase assortment—such signals are easier to receive and more effectively discriminate similarity. The churlish receiver strategy, $\alpha = 1$, is favored by this context, because it also increases assortment with similar individuals. Therefore this context is hostile to covert signals and to the receiver strategy that favors them.

To support the above statements, we demonstrate that the gradient for p in this context is always negative and that the gradient for α in this context is always positive. The gradients in this context are:

$$\begin{aligned} g(p)|_{q=1} = {} & (-1)\frac{s(1-s)}{Z^2}\big(1-(\alpha+(1-\alpha)(1-p)R)v_{\mathrm{D}}\big) \\ & \big((1-\alpha v_{\mathrm{D}})+(R(1-p)+pr)(\alpha v_{\mathrm{D}}+v_{\mathrm{L}})\big) \\ & \big(R(1-\alpha v_{\mathrm{D}})(v_{\mathrm{D}}+v_{\mathrm{L}})-r(1-v_{\mathrm{D}}(\alpha(1-R)+R))(\alpha v_{\mathrm{D}}+v_{\mathrm{L}})\big) \qquad (10) \\ g(\alpha)|_{q=1} = {} & \frac{s(1-s)}{Z^2}(1-w_{\mathrm{D}})\big(1-(1-w_{\mathrm{D}})(\alpha+(1-\alpha)(1-p)R)\big) \\ & \big((1-p)R(w_{\mathrm{L}}-w_{\mathrm{D}})(1-pr-(1-p)R)+prw_{\mathrm{L}}\big) \\ & \big((1-(1-p)R-pr)(1-\alpha(1-w_{\mathrm{D}}))+(R(1-p)+pr)w_{\mathrm{L}}\big) \qquad (11) \end{aligned}$$

where Z is a complex normalizing term (analogous to the partition function in statistical mechanics) and the symbols $v_{\mathrm{D}} = 1 - w_{\mathrm{D}}$ and $v_{\mathrm{L}} = w_{\mathrm{L}} - 1$ are used for compactness of notation. By inspection, every term after the leading (-1) in $g(p)$ is positive, for all allowed values of the variables, and so the gradient is always negative. Similarly, every term in $g(\alpha)$ is positive, and so the gradient is always positive. Therefore payoff incentives in the free choice context never favor covert signals and always favor churlish receivers.

While this context always favors overt signalers and churlish receivers, the strength of the incentives may vary. First, both gradients are proportional to the variance in similarity, $s(1-s)$. This indicates that intermediate similarity more strongly favors overt signals, unlike the situation in the forced choice context, in which high similarity favored overt signals. The reason that the variance matters now is that payoffs depend directly upon similarity, not upon attitudes. The more variance in similarity in the population, the greater the advantage of efficient assortment.

Overt signals have the advantage in this context, because they are better at assortment. Therefore, any change in variables that reduces the accuracy of assortment overall will reduce overt signalers' advantage. The important variables are R, the probability an overt signal is received, and the assortment proportional odds w_{D} and w_{L}. Reducing R reduces overt signalers' advantage, because it makes signals less valuable overall. Making either w_{D} or w_{L} closer to 1 makes assortment, based on attitudes, less effective. This also reduces overt signalers' advantage.

Joint dynamics: When do covert signals evolve? When both contexts matter, the joint dynamics take on one of three characteristic regimes. First, covert signals both invade and are evolutionarily stable. Second, overt signals both invade and are evolutionary stable. Third, a mixed equilibrium exists at which covert and overt signals coexist in the population. Figure 2 illustrates these three regimes.

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

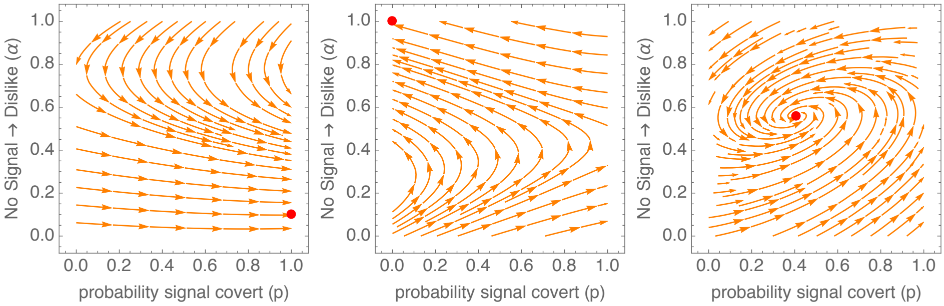

Figure 2. The three dynamic regimes that arise from the joint dynamics. In each plot, the paths show the evolutionary trajectories in each region of the phase space defined by the probability of covert signals (p, horizontal) and the probability of churlish receivers (α, vertical). The red points show equilibria. In all three plots: $d = 0.1$, $\delta = 0.01$, $q = 0.5$. Left: $s = 0.1$, $r/R = 0.25$, $w_{\text{L}} = 1.1$, $w_{\text{D}} = 0.9$. Middle: Same as left, but $w_{\text{D}} = 0.6$. Right: $s = 0.2$, $r/R = 0.75$, $w_{\text{L}} = 1.25$, $w_{\text{D}} = 0.8$.

These three examples all weigh the forced choice and free choice contexts equally, $q = 0.5$. The other parameters then shift the strength of incentives in each context to influence overall dynamics. For other values of q, the strength of incentives would have to shift as well to overcome weight given to each context.

When covert signals are sufficiently noisy (r/R low), similarity is sufficiently rare (s low), and assortment (w_{L}, w_{D}) not too efficient, covert signals can both invade and are an ESS. This situation is shown in the lefthand plot. While dynamics do not favor covert signals when α is large, near the top of the phase space, dynamics in that region favor smaller values of α. Eventually, α becomes small enough to allow covert signals to invade and reach fixation. In many cases, as in this one, a small amount of churlish receiver strategy, $\alpha > 0$, persists.

When the conditions outlined above are not met, incentives favor instead overt signals. The middle plot illustrates a case essentially the opposite of the one on the left. Here, $w_{\text{D}} = 0.6$, making assortment efficient. When assortment is efficient, recall, it may pay to dislike by default. This sets up a dynamic that eventually favors overt signals. While covert signals are still favored when α is low, the fact that larger values of α are favored everywhere leads eventually to invasion and fixation of overt signals.

Finally, the plot on the right shows an intermediate case, in which conditions favor both signaling strategies. Here $s = 0.2$, $r/R = 0.75$, $w_{\text{L}} = 1.25$, and $w_{\text{D}} = 0.8$. In this regime, the conditions that favor covert signals also favor more churlish receivers. Similarly, the conditions that favor overt signals also favor fewer churlish receivers. In total, the population comes to rest with a mixture of signaling and receiving strategies.

A more general view of the dynamics is available by considering the boundary conditions that make covert signaling an ESS. Recall that q is the relative importance of free choice scenarios. Define a threshold $\hat{q}$ as the largest value of q for which covert signals can resist invasion by overt signals. This is defined by values of $q = \hat{q}$

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

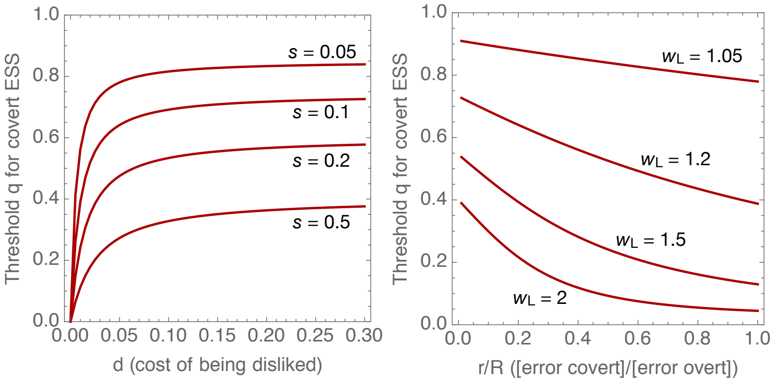

FIGURE 3. Plots of the largest value of q, $\hat{q}$, that allows covert signaling to be an ESS. Each curve represents a set of parameter values. Points below each curve make covert signals uninvadvable by overt signals. Points above each curve allow overt signaling to invade. Left: The cost of being disliked, d, for four values of the baseline rate of similarity, s. $r/R = 0.5$, $w_{\text{L}} = 1.1$, $w_{\text{D}} = 0.9$, and $\delta = 0.01$. Right: The ratio of covert transmission error, r, to the rate of overt transmission error, R, for four values of $w_{\text{L}} = 1/w_{\text{D}}$. $s = 0.1$, $R = 0.5$, $d = 0.1$, and $\delta = 0.01$.

and $\alpha = \hat{\alpha}$ that satisfy $g(p)|_{p=1} = 0$ and $g(\alpha)|_{p=1} = 0$. These cannot in general be solved analytically. But we can and do solve the system numerically, in order to illustrate the range of joint dynamics. Values of q less than $\hat{q}$ make covert signals evolutionarily stable—overt signals cannot invade. Values of q greater than $\hat{q}$ allow overt signals to invade, though we note that covert signals may nevertheless remain in the population, at an internal stable value p. Therefore $\hat{q}$ provides a useful metric of how strongly a parameter configuration favors covert signals.

We use $\hat{q}$ to summarize the tradeoffs in the signaling model. Recall that the cost of being disliked, d, is needed to favor covert signals. Therefore increasing d makes it easier for covert signals to be an ESS. However, the rate of similarity, s, favors overt signals. It is of value to note that d cannot compensate for s—if s is large, then steeply increasing costs d will not favor covert signals. We show this relationship in Figure 3, lefthand plot. Each curve in this plot is a threshold $\hat{q}$, below which covert signaling is an ESS. For each level of s, the impact of increasing d diminishes rapidly. Therefore some cost d is necessary for covert signals to evolve and be stable, but these costs cannot easily compensate when similarity is sufficiently common.

Consider another important pair of dimensions: the ratio of error rates r/R in covert/overt signals and the efficiency of assortment, as measured by w_{L} and w_{D}. Figure 3, righthand side, shows $\hat{q}$ curves for four values of $w_{\text{L}} = 1/w_{\text{D}}$, as functions of r/R. Covert signals are favored when r/R is small, as explained in the previous sections. But when assortment is very efficient, such as $w_{\text{L}} = 2$ near the bottom of the plot, it requires very low values of r/R to compensate in favor of overt signaling.

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

The dynamics of cooperation are more complicated than implied by models in which maximal benefits accrue to those who can simply avoid free riders. Not all cooperators are equal. Individuals vary, making assortment among cooperators important. Circumstances also vary. When individuals must occasionally collaborate with those outside their circles of friends, it can be critical to avoid burning bridges with dissimilar members of one's group. Covert signaling makes this possible, and this may be why it is sometimes observed in human societies, at both small and large scales.

We have shown that covert signaling is favored when forced choice scenarios are common, when similarity is low, when the cost of being disliked is high, and when covert signals are sufficiently noisy to make the meaning of a signal's absence ambiguous. We emphasize that covert signaling can be favored even though it is less effective than overt signaling at communicating similarity, because it simultaneously avoids communicating dissimilarity.

Our model also points to interesting transitions from inter- to intra-group assortment dynamics. As noted above, overt signaling systems are favored when the ability to assort on attitudes is high, and when being disliked by dissimilar individuals carries little risk. This is precisely the kind of situation that is assumed to obtain in inter-group assortment, where overt signals such as ethnic markers are used to discriminate between similar and dissimilar individuals. In these between-group contexts, the difference between similar and dissimilar individuals is so great that attempting to coordinate with dissimilar others is not worth the effort, and one can afford to burn bridges with them in order to ensure that similar others are aware of their similarity (McElreath et al., 2003). In fact, it might be argued that burning bridges with dissimilar out-group members is as much a goal of overt signals like ethnic markers as is attracting similar in-group members.

Intra-group assortment, however, is not simply a matter of scaling down inter-group dynamics. In this case, we must already presume some baseline level of similarity resulting from inter-group assortment; for there to be a group within which to assort, some degree of similarity should already be in place that defines that group, such as the shared interaction norms, communication systems, etc. that ethnic markers are thought to ensure. The benefits of further assorting on the basis of more nuanced similarity are therefore likely to be marginal relative to random assortment within the group. When the benefits to assortment on similarity are very strong, overt signaling is almost always favored, in part because they bring us out of the domain of intra-group assortment to that of inter-group assortment. When those benefits are less dramatic, however, the costs of burning bridges with dissimilar group members make covert signaling worthwhile.

Relatedly, we emphasize that the probability of similarity, s, need not reflect some number of discrete types in the population, but can instead refer to a level of selectivity in how much a given pair of individuals needs to have in common in order to be considered "similar." That is, s refers to the proportion of the population that would be considered similar to a focal individual in a given context, with higher values indicating a looser concept of "similar" than lower ones (e.g., $s = 0.7$ indicates that the individuals are choosing partners on the basis of whatever criteria would include the 70% of the group that are most similar to them). Changes to

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

s can have a significant impact on the overall dynamics of the system, because s essentially *defines* the concept of "similarity." When $s = 1$, 100% of individuals within the group will be considered "similar," and the only useful signals will be overt ethnic markers that allow individuals to avoid dissimilar out-group members. As s decreases, the criteria for considering a potential partner sufficiently similar to reap the benefits of enhanced coordination become stricter, as individuals deem a smaller proportion of the group worthy of homophilic assortment. The utility of covert signals increases as people are forced to be more choosy.

An interesting direction for future exploration is how these dynamics might respond to increased social complexity. In the larger and more complex societies associated with the development of agriculture, and particularly in the last few centuries, interactions with strangers have been increasingly frequent, necessitating strategies for temporary assortment (Johnson & Earle, 2000; Smaldino, 2016). In large, diverse populations, highly similar individuals should be rare, while the need for large scale cooperation in collective endeavors such as warfare, politics, or commerce would make it costly to burn bridges with these variously dissimilar partners. It is therefore likely that increases in social complexity would select for more complex covert signaling strategies.

For example, within complex industrialized societies, individuals often use Gestalt descriptions connoting a suite of information about the sort of person they are, which we call "social identities." Identity signaling, whether through overt social markers or through more covert communication, can be used by individuals looking to find others similar to themselves and to avoid being mistaken for something they are not (Berger & Heath, 2008; Smaldino, 2016). If the need to cooperate with dissimilar individuals is unlikely or if similar individuals are common, then overt declarations of identity should be expected. On the other hand, if burning bridges is both costly and likely given an overt signaling strategy, we should expect the relevant identity to be signaled much more subtly. There may be layers to how identity is signaled, with increasing levels of specificity signaled in increasingly covert ways, and without all received signals actively inducing a disposition of either liking or disliking toward the sender. A related signaling strategy, not covered by our model, might facilitate liking between similar individuals but only indifference otherwise. Using these "semi-covert" signals, individuals would be aware of failures to match, but simply not care. Casual, coarse-grain identity signaling may often take this form, as in cases of fashion adoption or pop culture allegiances. It would be interesting to investigate how common these kind of semi-covert signals are in small-scale communities, as they seem pervasive in complex industrialized societies.

In addition, our model helps make sense of findings from political psychology suggesting that people in the industrialized West who identify as conservative or right-leaning tend to view ambiguous people as hostile, while those identifying as liberal or left-leaning tend view ambiguous people as neutral (Vigil, 2010; Hibbing et al., 2014; Holbrook et al., 2016). In our model, a default attitude to dislike was strongly linked with overt signaling, which we in turn associate with the preservation of strong between-group boundaries. In contrast, a default attitude of neutral was associated with covert signaling, and with the avoidance of burned bridges to facilitate more widespread within-group cooperation. As a broad generalization, our analysis suggests that conservatives may be operating under the assumptions

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of stronger ingroup/outgroup boundaries, increased expectations of similarity toward those they signal, and lower costs to being disliked by dissimilar individuals. In contrast, liberals may be operating under the assumptions of a more broadly defined ingroup, limited expectations for similarity toward those they signal, and higher costs to being disliked by dissimilar individuals. Lending modest support to this idea is the finding that conservatives appear to have a stronger "need for cognitive closure" (reviewed in Hibbing et al., 2014), which is associated with, among other things, a distaste for uncertainty and ambiguity. The modeling framework we present in this paper may thus be useful in understanding patterns of differences between groups, including but not limited to political affiliation.

Ours is the first model of covert signaling. As such, it necessarily involves simplifying assumptions concerning the nature of signaling and cooperative assortment. For example, while we have allowed for covert signaling errors in the form of failed transmission to similar individuals, we have not included the converse form of error, where dissimilar individuals *are* able to detect the signal some of the time, and therefore update their disposition to disliking the covert signaler. Adding an additional parameter to account for this possibility does not qualitatively change our analysis. But it may create conditions where a non-signaling "quiet" strategy could invade. In addition, we ignore the possibility of strategic action on the part of the receiver to either improve coordination or to avoid partnering with dissimilar individuals entirely. We assumed that a pairing of dissimilar partners would simply lead to an unsuccessful collaboration, but such a pairing might instead lead each individual to pursue more individualistic interests. At the population level, we assumed that all individuals had an equal probability of encountering similar individuals, and that all similar and dissimilar individuals were equivalent. In reality, some individuals may be more or less likely to encounter similar individuals, perhaps related to differences in the tendency to be conformity- versus distinctiveness-seeking (Smaldino & Epstein, 2015), or reflecting minority-majority dynamics (Wimmer, 2013). Exploration of this variation opens the door to evaluating signaling and assortment strategies in stratified groups. All of these limitations provide avenues for future research that build upon the central findings reported here.

In a population where individuals vary and burning bridges is costly, overtly announcing precisely where one stands entails venturing into a zone of danger. Covert signaling, as in the case of humor or otherwise encrypted language, allows individuals to effectively assort when possible while avoiding burned bridges when the situation calls for partnerships of necessity.

Acknowledgments

Thanks to Patrick Barclay, William Baum and John Bunce for helpful comments on an early version of this paper. This work was supported by National Science Foundation Grant BCS 1357240 (to T.F. and R.M.) and by the Division of Social Sciences Dean's Office at the University of California Davis.

References

Antal, T., Ohtsuki, H., Wakeley, J., Taylor, P. D., & Nowak, M. A. (2009). Evolution of cooperation by phenotypic similarity. *Proceedings of the National Academy of Sciences*, *106*(21), 8597–8600.

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Aragonès, E. & Neeman, Z. (2000). Strategic ambiguity in electoral competition. *Journal of Theoretical Politics*, *12*(2), 183–204.

Barth, F. (1969). Introduction. In F. Barth (Ed.), *Ethnic Groups and Boundaries* (pp. 9–38). New York: Little, Brown.

Berger, J. & Heath, C. (2008). Who drives divergence? identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *Journal of Personality and Social Psychology*, *95*(3), 593.

Boyd, R. & Richerson, P. J. (1994). The evolution of norms: An anthropological view. *Journal of Institutional and Theoretical Economics*, *150*(1), 72–87.

Calcott, B. (2008). The other cooperation problem: Generating benefit. *Biology and Philosophy*, *23*(2), 179–203.

Castro, L. & Toro, M. A. (2007). Mutual benefit cooperation and ethnic cultural diversity. *Theoretical Population Biology*, *71*(3), 392–399.

Chwe, M. S.-Y. (2001). *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton University Press.

Cohen, E. & Haun, D. (2013). The development of tag-based cooperation via a socially acquired trait. *Evolution and Human Behavior*, *34*(3), 230–235.

Efferson, C., Lalive, R., & Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism. *Science*, *321*(5897), 1844–1849.

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. *Communication Monographs*, *51*(3), 227–242.

Flamson, T. & Barrett, H. (2008). The encryption theory of humor: A knowledge-based mechanism of honest signaling. *Journal of Evolutionary Psychology*, *6*(4), 261–281.

Flamson, T. J. & Bryant, G. A. (2013). Signals of humor: Encryption and laughter in social interaction. In M. Dynel (Ed.), *Developments in Linguistic Humour Theory*, volume 1 (pp. 49–73). Amsterdam: John Benjamins Publishing.

Gersick, A. & Kurzban, R. (2014). Covert sexual signaling: Human flirtation and implications for other social species. *Evolutionary Psychology*, *12*(3), 549–569.

Hammond, R. A. & Axelrod, R. (2006). The evolution of ethnocentrism. *Journal of Conflict Resolution*, *50*(6), 926–936.

Hibbing, J. R., Smith, K. B., & Alford, J. R. (2014). Differences in negativity bias underlie variations in political ideology. *Behavioral and Brain Sciences*, *37*, 297–350.

Holbrook, C., López-Rodríguez, L., Fessler, D. M. T., Vázquez, A., & Gómez, A. (2016). Gulliver's politics: Conservatives envision potential enemies as readily vanquished and physically small. *Social Psychological and Personality Science*.

Johnson, A. W. & Earle, T. K. (2000). *The Evolution of Human Societies: From Foraging Group to Agrarian State*. Stanford University Press.

López, I. H. (2014). *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism & Wrecked the Middle Class*. Oxford University Press.

Luce, R. D. & Raiffa, H. (1957). *Games and Decisions*. Wiley.

Mace, R. & Holden, C. J. (2005). A phylogenetic approach to cultural evolution. *Trends in Ecology and Evolution*, *20*(3), 116–121.

McElreath, R., Boyd, R., & Richerson, P. J. (2003). Shared norms and the evolution of ethnic markers. *Current Anthropology*, *44*(1), 122–130.

Moffett, M. W. (2013). Human identity and the evolution of societies. *Human Nature*, *24*(3), 219–267.

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Nettle, D. & Dunbar, R. (1997). Social markers and the evolution of reciprocal exchange. *Current Anthropology*, *38*, 93–99.

Ostrom, E. (2000). Collective action and the evolution of social norms. *The Journal of Economic Perspectives*, *14*(3), 137–158.

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838.

Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard University Press.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.

Smaldino, P. E. (2014). The cultural evolution of emergent group-level traits. *Behavioral and Brain Sciences*, *37*, 243–295.

Smaldino, P. E. (2016). The evolution of the social self: Multidimensionality of social identity solves the coordination problems of a society. In A. C. Love & W. C. Wimsatt (Eds.), *Beyond the Meme: Dynamical Structures in Cultural Evolution*. University Minnesota Press.

Smaldino, P. E. & Epstein, J. M. (2015). Social conformity despite individual preferences for distinctiveness. *Royal Society Open Science*, *2*(3), 140437.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation. *Current Anthropology*, *53*(6), 673–692.

Vigil, J. M. (2010). Political leanings vary with facial expression processing and psychosocial functioning. *Group Processes & Intergroup Relations*, *13*(5), 547–558.

Wimmer, A. (2013). *Ethnic Boundary Making: Institutions, Power, Networks*. Oxford University Press.

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

APPENDIX (ONLINE SUPPLEMENT)
The Evolution of Covert Signaling

Appendix A. Model derivation

A.1. **Payoff expressions.** As explained in the main text, we can define a general payoff expression for an individual with strategy p' and attitude matrix $\mathbf{a}'$ in a population with strategy $\{p, \mathbf{a}\}$. This expression is:

$$W(p', \mathbf{a}') = \Omega + q \Pr(\text{similar}|p', \mathbf{a}') + \\ (1-q)\big(1 - \Pr(\text{Disliked}|p', \mathbf{a}')d - \Pr(\text{Disliked}|p', \mathbf{a}') \Pr(\text{DislikeOther}|p', \mathbf{a}')\delta\big) \tag{12}$$

where Ω is an expected baseline payoff due to other activities.

A.2. **Context 1 probabilities.** $\Pr(\text{similar}|p', \mathbf{a}')$ is the probability of ending up in a similar dyad, given the focal has signaling strategy $\{p', \mathbf{a}'\}$. By definition:

$$\Pr(\text{similar}|p', \mathbf{a}') = \Pr(\text{LL}|p', \mathbf{a}') \Pr(\text{similar}|\text{LL}, p', \mathbf{a}') + \Pr(\text{LN}|p', \mathbf{a}') \Pr(\text{similar}|\text{LN}, p', \mathbf{a}') \\ + \Pr(\text{NN}|p', \mathbf{a}') \Pr(\text{similar}|\text{NN}, p', \mathbf{a}')$$

The terms like $\Pr(\text{similar}|\text{NN}, p', \mathbf{a}')$ are defined by conditional probaility:

$$\Pr(\text{similar}|\text{NN}, p', \mathbf{a}') = \frac{\Pr(\text{similar}, \text{NN}|p', \mathbf{a}')}{\Pr(\text{NN}|p')}$$

Defining the two terms on the right requires defining probabilities for dyad formation.

The probability that an LL dyad forms is:

$$\Pr(\text{LL}|p', \mathbf{a}') = \frac{p(\text{LL}|p', \mathbf{a}')w_L^2}{Z}$$

where the denominator Z normalizes the probability and $p(\text{LL}|p', \mathbf{a}')$ is the raw proportion of dyads that are LL, post signaling. Under the baseline receiver strategy, $\underline{\mathbf{a}}'$, it is defined as:

$$p(\text{LL}|p', \underline{\mathbf{a}}') = s\big(p'pr^2 + (p'(1-p) + (1-p')p)rR + (1-p')(1-p)R^2\big)$$

While we later develop this expression in general for all receiver strategies, it's worth consider the specific expression above, for sake of comprehension. To understand this expression, consider that LL dyads must comprise similar individuals, and both signals have to be received for the attitudes to form. Individuals are similar s of the time. There are three ways this can happen: (1) both individuals signal covertly $p'p$ of the time, (2) one individual signals covertly and the other overtly $p'(1-p) + (1-p')p$ of the time, or (3) both individuals signal overtly $(1-p')(1-p)$ of the time. In all three cases, both signals must be received. Probabilities for each of the other dyad types—LN, NN, ND, and DD—are defined similarly. Further down, we define all of these probabilities in general, using a more algorithmic approach.

The denominator Z normalizes the probabilities of each dyad forming. It is merely the sum of all of the numerators in the probabilities of different types of dyads. These other probabilities are defined similarly:

$$\begin{aligned}
\Pr(\text{LN}|p', \mathbf{a}') &= p(\text{LN}|p', \mathbf{a}')w_L/Z \\
\Pr(\text{LD}|p', \mathbf{a}') &= p(\text{LD}|p', \mathbf{a}')w_L w_D/Z \\
\Pr(\text{NN}|p', \mathbf{a}') &= p(\text{NN}|p', \mathbf{a}')/Z \\
\Pr(\text{ND}|p', \mathbf{a}') &= p(\text{ND}|p', \mathbf{a}')w_D/Z \\
\Pr(\text{DD}|p', \mathbf{a}') &= p(\text{DD}|p', \mathbf{a}')w_D^2/Z
\end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A.3. **Building generalized probabilities.** Now we construct the general probabilities that comprise the payoff expression by using a table of interactions.

Focal	Other	Sim	Signals	Probability of dyad	[6]FL	[7]FN	[8]FD	[9]OL	[10]ON	[11]OD
C	C	1	11	$p'psr^2$	a'_{SL}	a'_{SN}	a'_{SD}	a_{SL}	a_{SN}	a_{SD}
C	C	1	10	$p'psr(1-r)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{SL}	a_{SN}	a_{SD}
C	C	1	01	$p'ps(1-r)r$	a'_{SL}	a'_{SN}	a'_{SD}	a_{NL}	a_{NN}	a_{ND}
C	C	1	00	$p'ps(1-r)^2$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
C	C	0	11	$p'p(1-s)r^2$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
C	C	0	10	$p'p(1-s)r(1-r)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
C	C	0	01	$p'p(1-s)(1-r)r$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
C	C	0	00	$p'p(1-s)(1-r)^2$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
C	O	1	11	$p'(1-p)srR$	a'_{SL}	a'_{SN}	a'_{SD}	a_{SL}	a_{SN}	a_{SD}
C	O	1	10	$p'(1-p)sr(1-R)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{SL}	a_{SN}	a_{SD}
C	O	1	01	$p'(1-p)s(1-r)R$	a'_{SL}	a'_{SN}	a'_{SD}	a_{NL}	a_{NN}	a_{ND}
C	O	1	00	$p'(1-p)s(1-r)(1-R)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
C	O	0	11	$p'(1-p)(1-s)rR$	a'_{DL}	a'_{DN}	a'_{DD}	a_{NL}	a_{NN}	a_{ND}
C	O	0	10	$p'(1-p)(1-s)r(1-R)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
C	O	0	01	$p'(1-p)(1-s)(1-r)R$	a'_{DL}	a'_{DN}	a'_{DD}	a_{NL}	a_{NN}	a_{ND}
C	O	0	00	$p'(1-p)(1-s)(1-r)(1-R)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
O	C	1	11	$(1-p')psRr$	a'_{SL}	a'_{SN}	a'_{SD}	a_{SL}	a_{SN}	a_{SD}
O	C	1	10	$(1-p')psR(1-r)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{SL}	a_{SN}	a_{SD}
O	C	1	01	$(1-p')ps(1-R)r$	a'_{SL}	a'_{SN}	a'_{SD}	a_{NL}	a_{NN}	a_{ND}
O	C	1	00	$(1-p')ps(1-R)(1-r)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
O	C	0	11	$(1-p')p(1-s)Rr$	a'_{NL}	a'_{NN}	a'_{ND}	a_{DL}	a_{DN}	a_{DD}
O	C	0	10	$(1-p')p(1-s)R(1-r)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{DL}	a_{DN}	a_{DD}
O	C	0	01	$(1-p')p(1-s)(1-R)r$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
O	C	0	00	$(1-p')p(1-s)(1-R)(1-r)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
O	O	1	11	$(1-p')(1-p)sR^2$	a'_{SL}	a'_{SN}	a'_{SD}	a_{SL}	a_{SN}	a_{SD}
O	O	1	10	$(1-p')(1-p)sR(1-R)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{SL}	a_{SN}	a_{SD}
O	O	1	01	$(1-p')(1-p)s(1-R)R$	a'_{SL}	a'_{SN}	a'_{SD}	a_{NL}	a_{NN}	a_{ND}
O	O	1	00	$(1-p')(1-p)s(1-R)^2$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}
O	O	0	11	$(1-p')(1-p)(1-s)R^2$	a'_{DL}	a'_{DN}	a'_{DD}	a_{DL}	a_{DN}	a_{DD}
O	O	0	10	$(1-p')(1-p)(1-s)R(1-R)$	a'_{NL}	a'_{NN}	a'_{ND}	a_{DL}	a_{DN}	a_{DD}
O	O	0	01	$(1-p')(1-p)(1-s)(1-R)R$	a'_{DL}	a'_{DN}	a'_{DD}	a_{NL}	a_{NN}	a_{ND}
O	O	0	00	$(1-p')(1-p)(1-s)(1-R)^2$	a'_{NL}	a'_{NN}	a'_{ND}	a_{NL}	a_{NN}	a_{ND}

The columns of this table specify:

(1) The focal individual's signaling strategy
(2) The other individual's signaling strategy
(3) Whether or not (0/1) the individual's are similar
(4) Whether or not (0/1) focal/other signals are received by the other individual in the pair. 11 indicates that both signals are received. 10 indicates that focal's signal was received, while other's signal was not.
(5) The probability of this pairing and pair of signal reception events
(6) The probability that the focal individual forms attitude L. When a signal is received from a similar individual, this is a'_{SL}. When no signal is received or a covert signal is received from a dissimilar individual, this is a'_{NL}. When an overt signal is received from a dissimilar individual, this is a'_{DL}.
(7) The probability that the focal individual forms attitude N
(8) The probability that the focal individual forms attitude D
(9) The probability that the other individual forms attitude L
(10) The probability that the other individual forms attitude N
(11) The probability that the other individual forms attitude D

Parameters marked by a prime, such as p' and a'_{SL}, indicate aspects of the focal individual's strategy, to be contrasted with population values. Again, we refer to the vector of attitude parameters with $\mathbf{a}$.

Call this table $\mathbf{M}$. To compute probabilities, we multiply specific terms in each row and then sum these products down the rows. For example, the probability that the focal individual and a random individual mutually like one another, $p(\text{LL}|p', \mathbf{a}')$, is defined by:

$$p(\text{LL}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,5}\mathbf{M}_{i,6}\mathbf{M}_{i,9} \tag{13}$$

bioRxiv preprint doi: https://doi.org/10.1101/132407; this version posted April 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The other probabilities are defined similarity:

$$p(\text{LN}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,5}(\mathbf{M}_{i,6}\mathbf{M}_{i,10} + \mathbf{M}_{i,7}\mathbf{M}_{i,9}) \tag{14}$$

$$p(\text{LD}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,5}(\mathbf{M}_{i,6}\mathbf{M}_{i,11} + \mathbf{M}_{i,8}\mathbf{M}_{i,9}) \tag{15}$$

$$p(\text{NN}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,5}\mathbf{M}_{i,7}\mathbf{M}_{i,10} \tag{16}$$

$$p(\text{ND}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,5}(\mathbf{M}_{i,7}\mathbf{M}_{i,11} + \mathbf{M}_{i,8}\mathbf{M}_{i,10}) \tag{17}$$

$$p(\text{DD}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,5}\mathbf{M}_{i,8}\mathbf{M}_{i,11} \tag{18}$$

To derive probabilities of similarity and attitudes, all that is required is to multiply each of the products above with the corresponding value in column 3. This defines:

$$\Pr(\text{sim}, \text{LL}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,3}\mathbf{M}_{i,5}\mathbf{M}_{i,6}\mathbf{M}_{i,9} \tag{19}$$

$$\Pr(\text{sim}, \text{LN}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,3}\mathbf{M}_{i,5}(\mathbf{M}_{i,6}\mathbf{M}_{i,10} + \mathbf{M}_{i,7}\mathbf{M}_{i,9}) \tag{20}$$

$$\Pr(\text{sim}, \text{LD}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,3}\mathbf{M}_{i,5}(\mathbf{M}_{i,6}\mathbf{M}_{i,11} + \mathbf{M}_{i,8}\mathbf{M}_{i,9}) \tag{21}$$

$$\Pr(\text{sim}, \text{NN}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,3}\mathbf{M}_{i,5}\mathbf{M}_{i,7}\mathbf{M}_{i,10} \tag{22}$$

$$\Pr(\text{sim}, \text{ND}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,3}\mathbf{M}_{i,5}(\mathbf{M}_{i,7}\mathbf{M}_{i,11} + \mathbf{M}_{i,8}\mathbf{M}_{i,10}) \tag{23}$$

$$\Pr(\text{sim}, \text{DD}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,3}\mathbf{M}_{i,5}\mathbf{M}_{i,8}\mathbf{M}_{i,11} \tag{24}$$

All that remains are probabilities that any random individual Likes or Dislikes the focal:

$$\Pr_F(\text{L}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,5}\mathbf{M}_{i,9} \tag{25}$$

$$\Pr_F(\text{D}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,5}\mathbf{M}_{i,11} \tag{26}$$

as well as the probabilities that the focal likes or dislikes the other individual:

$$\Pr_O(\text{L}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,5}\mathbf{M}_{i,6} \tag{27}$$

$$\Pr_O(\text{D}|p', \mathbf{a}') = \textstyle\sum_{i=1}^{32} \mathbf{M}_{i,5}\mathbf{M}_{i,8} \tag{28}$$