AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

© 2021 American Psychological Association
ISSN: 0033-295X

https://doi.org/10.1037/rev0000344

Covert Signaling Is an Adaptive Communication Strategy in Diverse Populations

Paul E. Smaldino and Matthew A. Turner
Department of Cognitive and Information Sciences, University of California, Merced

Identity signals are those common components of communication transmissions that inform receivers of the signaler's membership in a categorizable subset of individuals. Such signals may be overt, broadcast to all possible receivers, or covert, encrypted so that only similar receivers are likely to perceive their identity-relevant meaning. Here we present an instrumental theory of covert signaling, based on the function of identity signals in social assortment. We argue that covert signaling is favored when signalers are generous toward strangers, when costs of being discovered as dissimilar are high, and when the ability to assort only with preferred partners is restricted. We further argue that covert signaling should be more common among members of "invisible" minorities, who are less likely to encounter similar individuals by chance. We formalize this theory with an evolutionary model to more rigorously explore the consequences of our assumptions. Our results have implications for our understanding of numerous aspects of social life, including communication, cooperation, social identity, humor, pragmatics, politics, hate speech, and the maintenance of diversity.

Keywords: signaling, coordination, cultural evolution, dog whistles, agent-based model

Much of communication seems to violate the cooperative principle that speech should be informative, relevant, and clear (Grice, 1975). For example, on the morning of November 9, 2016, a U.S.-based Twitter user cryptically tweeted, "Remember that scene in 'The Two Towers' where Theoden is watching the Urak Hi surround Helms Deep?" The meaning of this public comment can appear baffling unless you (a) are aware of the film (or book) in question as the second installment of *The Lord of the Rings* trilogy and know that the scene involves characters being threatened with destruction by an evil and inhuman horde, (b) can relate to a sentiment of catastrophic defeat felt by many left-leaning and moderate U.S. voters on the day after the 2016 presidential election, and (c) understand that sentiment as the appropriate contextual backdrop for the comment. In other words, the tweet's meaning is unclear unless you share (or at least are sympathetic to) aspects of the speaker's social identity. It is certainly much less direct that saying "I am unhappy because Trump won the election." Encrypted signals of identity like this are extremely common, but their function is still not well understood. Our purpose in this article is to explain why it may be advantageous to use covert or encrypted expressions of identity instead of more overt declarations, and, of equal importance, to delineate the sociocultural conditions under which that advantage will or will not hold.

Identity signals are those components of communication transmissions, displays, or other behaviors that inform receivers of the signaler's membership (or, minimally, nonmembership) in some categorizable subset of individuals. Identity signals are a core part of human social life in large and diverse populations. They allow us to make rapid decisions about potential social partners, to assort with desirable partners, and avoid undesirable ones (Berger & Heath, 2008; Hutchison & Martin, 2015; Miller & Todd, 1998; Smaldino, 2019). It has been argued that identity, and the associated ability to signal similarity and thereby facilitate cooperation between strangers, is foundational to the evolution of complex societies (Moffett, 2019; Smaldino, 2019).

Covert identity signals are purposely obscure, and their significance is often missed even by their intended audience. It may appear counterintuitive that the use of covert signals could be adaptive because maximizing clarity often seems like the highest pursuit of communication. Yet, the existence of covert signals suggests some adaptive function. What might this function be? What situations most strongly call for covert signaling? Among what sort of communities should we most expect covert signals? And what sort of societies do norms of covert signaling facilitate?

We answer these questions by developing an instrumental theory of identity signaling for cooperative assortment, building on previous work (Smaldino, 2019; Smaldino et al., 2018). We explain why, in a

Paul E. Smaldino https://orcid.org/0000-0002-7133-5620
Matthew A. Turner https://orcid.org/0000-0001-7421-4552

We thank Greg Bryant, Tom Flamson, and Tamara van der Does for helpful comments on an earlier draft, and Valerie Moss for assistance with our review of the literature. This article was improved by stimulating discussions with Mirta Galesic, Tamara van der Does, Zackary Dunivin, Michael Chwe, and audiences at Stanford, UC Davis, the University of Glasgow, and the Santa Fe Institute.

This work was supported by ARO Grant W911NF-20-1-0220 and by a Hellman Fellows Fund award to P.E.S. Model simulations were run on the MERCED computing cluster hosted at UC Merced, which is supported by the National Science Foundation (Grant No. ACI-1429783). An earlier version of this article was posted as a preprint on *SocArXiv*.

Correspondence concerning this article should be addressed to Paul E. Smaldino, Department of Cognitive and Information Sciences, University of California, Merced, 5200 Lake Road, Merced, CA 995343, United States. Email: psmaldino@ucmerced.edu

diverse population, covert signaling will often be the preferred method of communicating identity information whenever individuals are sufficiently likely to interact with dissimilar partners and when having one's identity revealed to those partners is sufficiently costly. Using agent-based models, we reach several conclusions about when we should expect identity signaling to be more covert, including the expectation that covert signals should be more common in diverse populations and among minority groups, and that these conditions will also select for more generous expectations concerning interactions with strangers.

Identity Signals Are for Cooperative Assortment

Humans may be the cooperative species (Bowles & Gintis, 2013), but not all cooperators are equal. Partner choice matters beyond the avoidance of free riders. Two potential partners might be equally inclined toward cooperation, but may differ in their goodness of fit for the particulars of the partnership. Important considerations in choosing cooperative partners include establishing common ground for communication (Clark & Brennan, 1991), the ability to establish shared intentions and a collective mindset (Tomasello et al., 2005; Gallotti & Frith, 2013; Moffett, 2019), and, more generally, having similar backgrounds, experiences, and traditions. In other words, it behooves individuals to coordinate on beliefs, norms, and goals.

In groups that are sufficiently insular and tight-knit, many of these similarities can be taken for granted. However, in larger populations where at least occasional interaction with strangers is required, signals for assortment become necessary (Moffett, 2019; Smaldino, 2019). Individuals regularly signal their identity, consciously and unconsciously, to find similar others or to reaffirm established ties. Such subsets can be large and reflect strong social boundaries, such as "Québécois" or "Muslim," or small and reflect subtler intragroup variation, such as "someone with relaxed views on monogamous sexual relationships." Identity signaling serves a key social function by enabling individuals to rapidly characterize others as similar or dissimilar. Finding similar others has many proximate psychological benefits, such as better mental health, or the security that results from a stronger sense of group identity. Ultimately, however, identity signaling serves to facilitate social assortment: preferentially interacting with similar individuals and avoiding the costs of interacting with dissimilar individuals. Of course, there are also benefits to be gained from diversity and division of labor rather than homogeneity. Even in these cases, however, we would argue that *some* minimum degree of similarity is beneficial to facilitate coordination on norms and goals. It is this type of assortment that is our focus.

Finding similar others is critical to maximize the benefits of cooperation. The importance of collaborative networks is hard to overstate: The very foundation of human sociality rests on the synergistic benefits of cooperation. Individuals working together can achieve much more than one person could alone, from raising children together, to hunting together to increase production, to working together to increase the efficacy of a political movement. Most social and evolutionary research on cooperation has focused on how cooperative behaviors can spread and resist exploitation by individuals who do not contribute anything. However, once the free-rider issue is resolved for a community, its members are still left with the *other* problem of cooperation: How to best generate a benefit between two or more cooperators (Calcott, 2008; Smaldino, 2014). The trick is not just to find *someone* cooperative, but to find the *best* person to cooperate with. This involves coordination. Individuals must find common ground to effectively communicate. Similar backgrounds, experiences, and cultural traditions help individuals to coordinate interests, as does the capacity to share intentions and adopt a group mindset[1]. Poor matches, in contrast, can invite costly conflict. These problems are solved, at least in part, by signals of social identity (Smaldino, 2019).

Social scientists have for a long time discussed how identity is used as a signal (Akerlof & Kranton, 2000; Barth, 1969; Donath, 1999; Goffman, 1978; Iannaccone, 1992), and an extensive formal literature has explored how arbitrary signals can facilitate assortment on norms and preferences (Castro & Toro, 2007; Efferson et al., 2008; McElreath et al., 2003; Nettle & Dunbar, 1997). Language can serve as a marker for social coordination (Nettle & Dunbar, 1997), as can visible purchasing and fashion choices (Berger & Heath, 2007, 2008; Brooks & Wilson, 2015; Smaldino et al., 2017). The logic works as follows. Individuals vary on important characteristics, and would like to be able to assort with others who share their characteristics. However, those characteristics are often opaque, making the assessment of potential partners difficult. A potential solution is the use of overt or conspicuous displays—often called tags or ethnic markers—that are perceivable by a broad audience. These displays include fashion, language, accents, and behavioral indicators (Barth, 1969; McElreath et al., 2003; Wimmer, 2013). Ethnic markers allow individuals to assess others' key identities and thereby assort with similar partners, providing the resulting pairs with the benefits of effective coordination. A large sociological literature has also documented the importance of ethnic markers in establishing and maintaining cooperative networks in multicultural settings in contemporary industrialized societies (Gold, 1991; Leong et al., 2020; Nee et al., 1994; Sanders, 2002; Wimmer, 2013).

By design, overt signals like ethnic markers lack ambiguity. They are visible to all, and most potential receivers will know what they mean. This creates a potentially serious problem. Overt signals facilitate assortment between group members, but they simultaneously exacerbate divisions between groups whenever highlighting differences burns bridges with out-group individuals. In other words, boundaries between in-group and out-group members become clearly delineated. There are surely cases where this delineation is both individually desirable and societally adaptive, as when the majority of interactions are expected to occur between members of the same group. Even in such cases, however, variation is not limited to the traits that delineate group boundaries. Individual variation on beliefs, norms, and goals exists *within* groups as well. This is readily apparent in the large, multicultural societies that comprise much of the developed world. Such variation creates a problem for overt signals of identity because burning bridges is not always an acceptable option. Dissimilar individuals may still need to cooperate with one another from time to time. This can occur in many contexts, from individuals in collective farming communities to coworkers in large companies to students working on group assignments to politicians crafting legislation. Highlighting differences between partners can then impede cooperation. In some cases, the identification of differences can be costly in the extreme, as in the cases of persecuted minority groups or

[1] Many cooperative endeavors benefit from diversity and division of labor. Even in these cases, however, a baseline of similarity on shared norms, goals, and expectations are required to achieve those benefits (Bicchieri, 2005).

political dissidents under authoritarian regimes. Under these conditions, an alternative to overt identity signaling is desirable.

A Theory of Covert Signaling

Building on prior work by Smaldino et al. (2018), we develop a theory of covert signaling. Covert signaling involves the transmission of information—and here we focus on information related to identity—that is accurately received by its intended audience but obscured when received by others. Implicit in this definition is the assumption of public communication—that is, of a broad audience with characteristics that are beyond the control of the signaler. The nature of covert signaling is well described by Loury (1994, p. 448) in the context of language: "If the significance of some words as signals of belief is known only to 'insiders,' their use in public allows the speaker to convey a reassuring message to some listener—'I share your values'—without alarming the others." Political dog whistling (Haney-López, 2015; Henderson & McCready, 2017) is probably the best-known example of this phenomenon, though as we will show, there are many others. This contrasts with overt signaling, in which honest signals of identity are widely broadcast.

We expect covert signaling to be common under conditions where individuals are likely to interact with others who differ in their beliefs, norms, and goals. More specifically, covert signaling functions to facilitate cooperative assortment in scenarios in which (a) individuals can use identity information to select cooperative partners who are similar, but (b) cooperative interactions must nevertheless occur at least sometimes between dissimilar group members, and (c) revelations of dissimilarity can further impede cooperation or lead to other costly outcomes. These conditions will be met when the ability to assort on similar characteristics, *homophily*, is weak, and when the risk of paying heavy costs when differences are revealed is high. We expect these conditions to be met in diverse populations, where differences are common, and among "invisible" minorities whose differences are socially important but whose identities are not obviously discernible from physical characteristics alone (including, though hardly limited to, LGBTQ+ individuals, religious minorities, and political dissidents).

Covert signals are presumed to be honest, not deceptive. The assumption of honesty applies to much communication, particularly signals that are widely broadcast and can therefore be scrutinized by a large audience. Honesty can be favored for costly signals that are difficult to fake (Grafen, 1990; Spence, 1973), but also when signals are low cost as long as individuals either are likely to interact repeatedly (Silk et al., 2000) or have interests in coordination (Farrell & Rabin, 1996). This article particularly concerns signals of the last variety. Because coordination implies that interests are aligned, individuals are incentivized to signal their identity honestly if those signals lead to optimal coordination. Nevertheless, honest signaling can also be costly if it incurs a negative response from audience members who are *not* optimal partners for coordination.

For this reason, honest signals need not be clear or direct. Indeed, the distinction between honest signaling and deception becomes blurry if one considers communication as feedback between individuals each trying to assess and manage their social environments, so that little communication is either purely honest or purely deceptive (Owings & Morton, 1997). There is often strategic value to being indirect ("Would you like to come up and look at my etchings?") or ambiguous ("We support academic freedom!"). Indirect speech can yield plausible deniability of intent when suggestions are rejected (Lee & Pinker, 2010). Ambiguous or vague speech allows for multiple interpretations by different receivers, maximizing apparent agreement and minimizing confident disagreement (Aragones & Neeman, 2000; Eisenberg, 1984; Yoon et al., 2020). Previous discussions of intentionally unclear communication are related to the present discussion, but these generally assume either a single receiver or that signals reflect particular static facts about the state of the world. In contrast, we focus on identity signaling for cooperative assortment, whereby the signal is broadcast to a wide and potentially diverse audience who have an interest in knowing the sort of person the signaler is. This is an instrumental interpretation of signaled information as bearing upon decisions concerning the formation of cooperative partnerships. Such partnerships could include friendships, romantic relationships, business arrangements, or those involving more transient interaction. The theory of covert signaling is predicated on the idea that communications often contain multiple layers of meaning, or implicatures, and that the receiver's background knowledge and perception of context affects whether and how those implicatures will be revealed (Clark & Schaefer, 1987; Grice, 1975; Searle, 1975; Sperber & Wilson, 1986).

Several lines of research provide evidence that people use covert signals of identity as we describe. Flamson and colleagues, using samples in both the urban U.S. and rural Brazil, examined humor as an encrypted signal of identity, and showed that having prior knowledge of a joke's subject matter increased participants' ratings of the joke as funny and that social closeness correlated strongly with similarity in rating jokes as funny (Flamson & Barrett, 2008, 2013; Flamson & Bryant, 2013). They propose that humorous utterances can serve as honest signals of similarity by implying multiple simultaneous meanings, so that perceiving something as funny requires decrypting meanings that are less obvious or superficial. Covert signals need not be humorous, however. Berger & Ward (2010) examined covert consumer choices in fashion. They found that fashion-savvy consumers prefer subtly, rather than overtly, marked products, as "they provide differentiation from the mainstream and should facilitate interaction with others 'in the know'" (Berger & Ward, 2010, p. 556). Fittingly, they found that fashion insiders not only preferred products without labels or with less well-known labels, but that preferences for such products increased in identity-relevant domains and in public settings where signals were likely to be observed by a diverse audience.

Our theory of covert signaling suggests that covert signals would be especially favored when the costs of having one's identity revealed by those who do not share that identity are potentially costly. There is some evidence for this in the gay community, historically a persecuted but "invisible" minority (and still quite persecuted in many parts of the world). Fischer (2015) documented gay men's fashion in San Francisco in the 1970s, detailing certain indicators that would reveal one's sexual orientation only to those who knew what to look for. In one of the few experimental investigations on the topic, Shelp (2003) found evidence suggesting that gay men were more accurate than straight men in identifying other gay men from merely watching muted video recordings. Relatedly, there has been much speculation, though to our knowledge little in the way of detailed investigations, into the use of covert signals by political dissidents under authoritarian regimes (Boyer, 2018; Kuran, 1989, 1995).

A key component of our theory of covert signaling is the importance of receiver strategies in relation to the question: How does a receiver assess a stranger in the absence of a clear identity signal? In a diverse society in which cooperation with strangers is important, we should expect receivers to be generous, and maintain openness or neutrality until more information can be gathered. When receivers are generous in this way, covert signals can be effectively utilized to keep the attitudes of dissimilar receivers neutral instead of negative. In contrast, receivers in a more insular or parochial society are expected to be churlish, maintaining negative attitudes toward strangers unless similarity to the in-group can be confirmed. In this case, covert signals may be less effective because honest signalers cannot avoid being disliked by dissimilar receivers and only weaken the strength of their signals to similar receivers. In our formal analysis, we do not assume that individuals are necessarily generous or churlish. Rather, we allow receiving strategies to coevolve with signaling strategies. In the next sections, we present and analyze a formal model to add precision to our theory. In our subsequent discussion, we will reflect on the social forces currently in play that shape the current selection pressures on signaling and receiving strategies, and the social divisions they potentially imply.

Model Description

Our agent-based model extends a simpler mathematical model analyzed previously by Smaldino et al. (2018). That model used replicator dynamics and evolutionary stability analyses—standard approaches in evolutionary game theory (McElreath & Boyd, 2007)—to consider the conditions under which a strategy of covert signaling could not be invaded by a rare overt signaler (invasion requires the rare variant to have a higher expected payoff against the dominant strategy). The agent-based instantiation presented here allows us to perform a number of analyses that are difficult or impossible with the simpler model, but which are important to a full exploration of the theory. We believe it is also a closer reflection of the verbal theory described above.

In accordance with best practices in describing complex agent-based models (Grimm et al., 2020), we first give an overview and brief description of the model, followed by a more detailed description. As with all such models, our model presents a highly simplified picture of reality. However, such simplifications allow for a much clearer exploration of the consequences that follow our assumptions (Levins, 1966; Smaldino, 2017, 2020).

In general, we assume the following:

- Individuals differ on a variety of characteristics, and may be considered similar if and only if they share some threshold number of characteristics in common. Similar pairs cooperate more effectively than dissimilar pairs.
- Individuals signal their identity either covertly or overtly. Covert signals are noisier and are only perceived by similar receivers. Receivers then form attitudes about signalers, which may be positive, negative, or neutral.
- Individuals pair up for cooperative tasks, using attitudes to influence partner selection. The extent to which homophilic assortment on attitudes is possible is an exogenous property of the social environment.
- Individuals receive payoffs from these cooperative partnerships. Payoffs are higher when pairs are similar, and lower when individuals hold negative attitudes toward one another.
- Individuals engage in success-biased copying, whereby the signaling (overt vs. covert) and receiving (generous vs. churlish) strategies of individuals with higher payoffs are preferentially copied.

In this way, we can explore how signaling and receiving strategies evolve in response to different properties of the social environment.

We consider a population in which N agents are characterized by a set of K fixed traits, representing features that matter for coordination, including personality traits, norms, beliefs, goals, and prior experiences. Each agent's traits are represented as a vector of binary values—an agent either does or does not possess any given trait, represented computationally as a 0 or 1. Two agents are *similar* if they share a sufficient fraction S of traits in common, otherwise they are dissimilar. Traits are not directly observable. Instead, agents signal information about themselves, as part of their normal quotidian behavior, which other agents can use to form attitudes about them. Signals are either *overt*, so that most agents are likely to receive them, or *covert*, in which case they are less easily perceived and then only by similar agents. Agents form attitudes toward other agents based on the signals of similarity or dissimilarity they receive. They like similar agents and dislike dissimilar agents. If an agent does not receive sufficient information to assess similarity, they may be *generous*, and remain neutral about the signaler, or they may be *churlish*, and dislike them in the absence of positive evidence of similarity. In summary, each agent is characterized by a vector of traits, a vector of attitudes toward each other agent, a signaling strategy, and a receiving strategy (Table 1). All global parameters are fully described in Table 2. Dynamics of the model proceed in discrete time steps, with each step consisting of three stages: (1) signal

Table 1
Agent and Interagent Attributes

Agent property	Description
a_{ij}	Attitude of agent i toward agent j; $\mathbf{a}_i$ is the N-dimensional attitude vector of agent i, where each item is in $\{-1, 0, 1\}$ and which correspond to attitudes of dislike, neutral, and like, respectively.
τ_{ik}	Trait k of agent i; τ_i is the K-dimensional trait vector of agent i, where each item is in $\{0, 1\}$.
S_{ij}	*Similarity* between agents i and j. Equal to $1 - h(\tau_i, \tau_j)$ where $h(\tau_i, \tau_j)$ is the Hamming distance between i and j's trait vectors.
Signaling strategy	Either *overt* or *covert*. Overt signaling always results in the communication of the signaler's traits, while covert signalers only reveal their traits to "similar" others; dissimilar others do not receive a covert signal.
Receiving strategy	Either *churlish* or *generous*. Churlish receivers default to dislike other agents from whom they have not received a signal. Generous receivers default to a neutral attitude toward others from whom they have not received a signal.

Table 2
Global Model Parameters

Symbol	Definition	Default value
N	Population size	100
w	*Homophily*, the ability of agents to preferentially interact with similar others	0.5
d	*Disliking penalty* imposed when one agent in interaction dyad dislikes the other	0.25
δ	*Synergistic disliking penalty* imposed when both interacting agents dislike each other	d
R	*Overt signaling efficiency*, the probability of receiving an overt signal	1.0
r	*Covert signaling efficiency*, the probability of receiving a covert signal	0.5
s	*Similarity benefit* received by similar interacting pair	0.25
K	Number of agent identity traits	9
S	Minimum fraction of traits in common for a pair to be considered similar	0.5
β	Selection strength for probability learner adopts teacher's strategy for a given difference between teacher and learner payoff.	10
M	Number of minority/majority traits used to define majority/minority populations.	$\frac{K-1}{2}$
n_R	Number of rounds of cooperative assortment per time step.	100
T	Number of time steps in each simulation run.	100

Note. In the final column, the default value for each parameter is indicated this value was used wherever another value is not explicitly specified.

transmission, (2) cooperative assortment, and (3) payoff-biased copying (see Figure 1).

Signal Transmission

Agents use one of two strategies to communicate information about themselves to others. Overt signalers send clear signals of their identity, which are received by a proportion R of the population. Anyone can perceive overt signals and use them to assess their similarity or lack thereof to the sender. Covert signalers send subtle or encrypted signals, which are less readily perceived and even then only by similar agents; a proportion $r \leq R$ of similar agents receive covert signals. Once signals are received, each agent forms an attitude toward every other agent of *like*, *dislike*, or *neutral*. All receivers like similar agents and dislike dissimilar agents. In cases where a signal was not received, generous receivers retain a neutral attitude toward the signaler while churlish receivers dislike them. Notice that covert signalers will be liked less often by similar receivers compared with overt signalers, but will never be disliked by dissimilar receivers who are generous.

Cooperative Assortment

After signals are sent and attitudes formed, agents engage in multiple rounds of cooperative assortment, in which each agent pairs up with exactly one other agent for a cooperative interaction. Attitudes aid assortment, so that in the search for interaction partners, agents can seek out partners they like and avoid partners they dislike. The extent to which such *homophily* is possible is a property of the social setting. Sometimes agents must cooperate with partners regardless of similarity or attitudes. The parameter w represents the strength of homophily. When homophily is perfect ($w = 1$), agents can always pair with those they like and avoid those they dislike. When homophily is nonexistent ($w = 0$), agents ignore attitudes and pair up at random. Intermediate values of homophily indicate partial but imperfect assortment based on attitudes. The algorithmic details of this process are provided in the Appendix.

Cooperative interactions yield payoffs. Similar pairs are better able to coordinate, due to sharing norms, goals, etc., and so receive larger payoffs. The baseline payoff for a cooperative interaction between dissimilar pairs is 1, with similar pairs receiving $1 + s$. Being disliked impedes cooperation because negative attitudes make it harder to interact. When one individual dislikes the other, she makes the interaction more difficult than it must be and thereby imposes a cost $-d$ on the pair's interaction. When both individuals dislike one another, their difficulties act synergistically, inducing an additional cost $-\delta$ on each. This cost could result from spite or from uncontrollable inefficiency, a negative consequence of second-order common knowledge (Chwe, 2001). For simplicity, we usually assume that $d = \delta$ (so that the penalty for mutual disliking was double that of one person disliking the other), though we vary this assumption in the Appendix. Agents undergo n_R rounds of cooperative assortment each time step, representing the fact that similarity and attitudes will inform many interactions over which payoffs may accumulate.

Payoff-Biased Copying

Agents change signaling and receiving strategies in a process of cultural evolution driven by success-biased transmission. That is, agents with higher payoffs are more likely to be copied. Each agent (the learner) observes another agent chosen at random and compares their payoffs. The learner decides to copy the observed agent with a probability that is determined by a sigmoid function that increases from 0.5 when the observed agent has a higher payoff and decreases from 0.5 when the observer has a higher payoff. More precisely, the probability that a learner i will adopt the strategy of an observed agent j is:

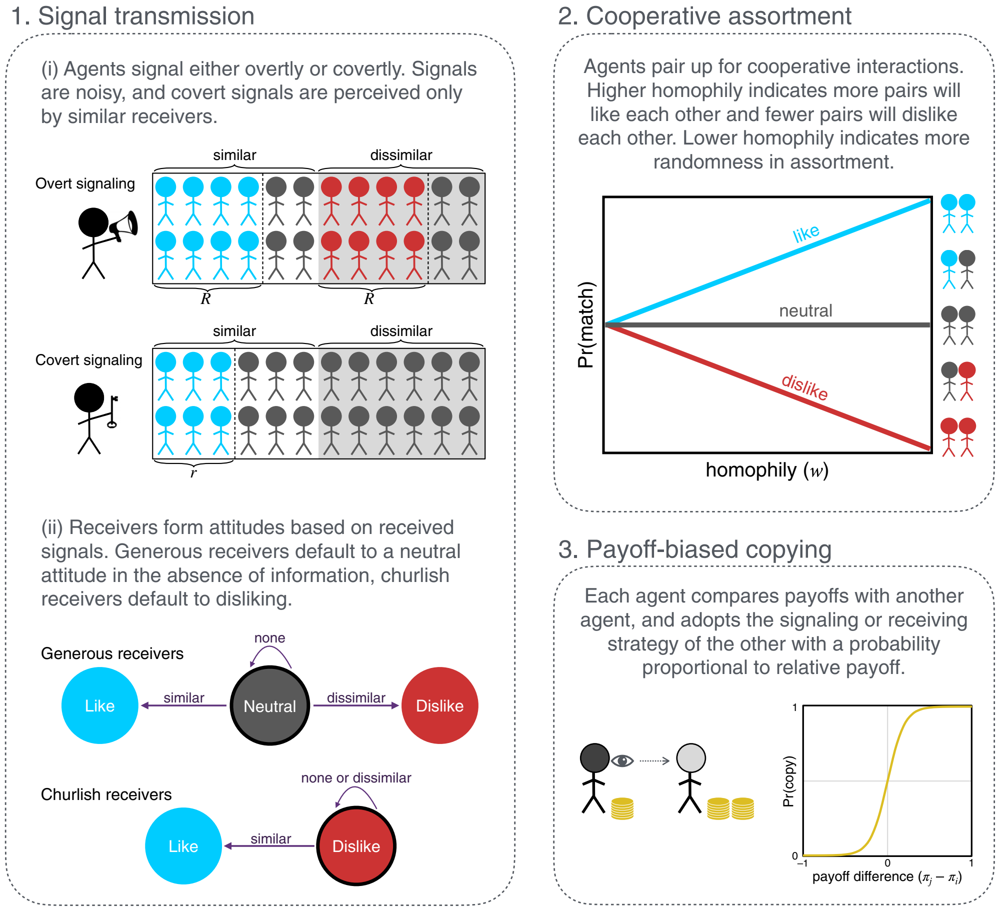
$$\Pr(i\,\text{copies}\,j) = \frac{1}{1 + e^{-\beta(\pi_j - \pi_i)}}, \tag{1}$$

where π_i is the total payoff of agent i accumulated over all interaction rounds, and β is a global parameter that specifies the level of stochasticity in this copying process. If copying occurs, the observer adopts either the signaling or the receiving strategy of the observed agent with equal probability.

Note that payoff-biased copying need not imply the transmission of any particular signals, which would need to be learned from specific individuals and would change over time in a process not modeled here. Rather, what is transmitted are the tendencies to signal in a covert versus overt way, and likewise to default to neutral or negative attitudes when acting as a receiver.

At the end of each time step, all payoffs and attitudes are reset. Model runs continue for $T = 100$ time steps, when equilibrium levels of signaling and receiving strategies are attained. Python code for this model can be found at https://github.com/mt-digital/identity-signaling.

Figure 1
Model Dynamics


1. Signal transmission
(i) Agents signal either overtly or covertly. Signals are noisy, and covert signals are perceived only by similar receivers.
Overt signaling
similar
dissimilar
R
R
Covert signaling
similar
dissimilar
r
(ii) Receivers form attitudes based on received signals. Generous receivers default to a neutral attitude in the absence of information, churlish receivers default to disliking.
Generous receivers
none
Like
similar
Neutral
dissimilar
Dislike
Churlish receivers
none or dissimilar
Like
similar
Dislike
2. Cooperative assortment
Agents pair up for cooperative interactions. Higher homophily indicates more pairs will like each other and fewer pairs will dislike each other. Lower homophily indicates more randomness in assortment.
Pr(match)
like
neutral
dislike
homophily (w)
3. Payoff-biased copying
Each agent compares payoffs with another agent, and adopts the signaling or receiving strategy of the other with a probability proportional to relative payoff.
Pr(copy)
1
0
−1
0
1
payoff difference (π_j − π_i)

Note. See the online article for the color version of this figure.

Results

Here, we present the results of our agent-based simulations. Unless otherwise indicated, all simulations are initialized with agents equally likely to be either overt or covert signalers and either generous or churlish receivers. All results represent the average across 100 simulations for each set of parameter values. Default parameter values are indicated in Table 2. In the Appendix, we perform robustness analyses, varying each of the model's many parameters. In general, we find that our results (presented below) are robust to a wide range of assumptions, offering strong support for the wide applicability of our conclusions.

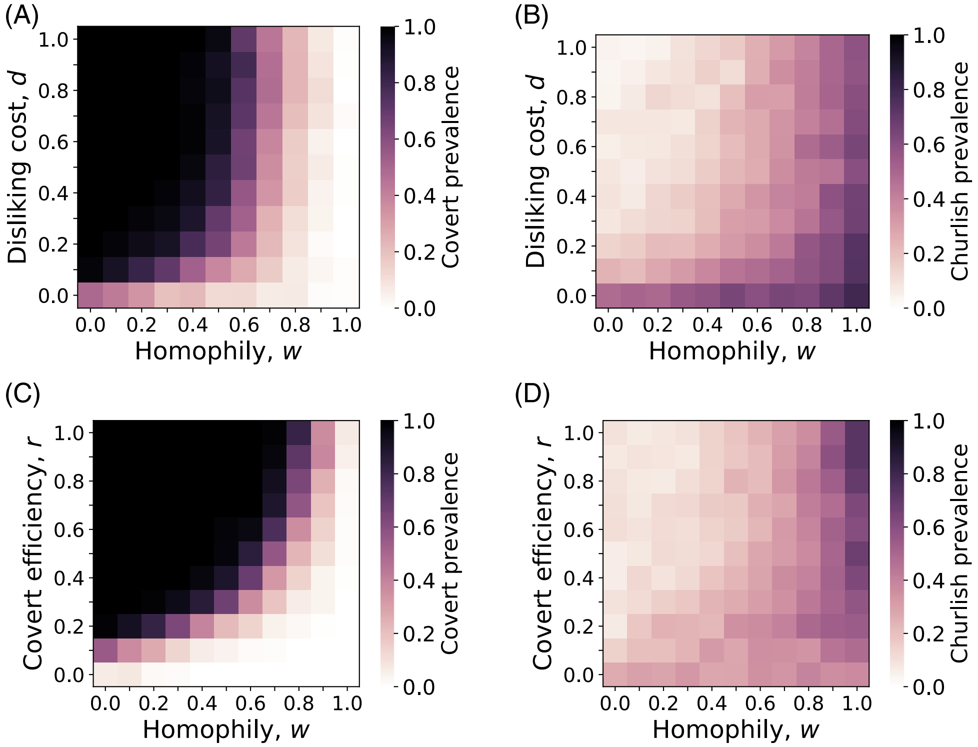
Covert Signaling Is Favored When Homophily Is Low, the Cost of Being Disliked Is High, and Covert Signals Are Efficient

In agreement with our theory and with previous analytical results (Smaldino et al., 2018), we find that covert signaling is favored when homophily is weak, particularly when the cost of disliking is sufficiently high and the efficiency of covert signals is sufficiently strong (Figure 2A, C). When homophily is strong, agents can easily use the attitudes they form from the signals they receive to assort with similar partners and avoid dissimilar partners. In such cases, overt signaling allows agents to more effectively sort similar from dissimilar others. When homophily is weak, however, agents must sometimes interact with dissimilar others. In this case, it is better to avoid having identified each other as dissimilar, and so covert signaling is favored. This result requires a sufficiently large cost of being disliked, *d*, so that the benefit of avoiding being identified as dissimilar outweighs the reduction in signal efficiency and the ability to avoid dissimilar partners altogether (Figure 2A). The evolution of covert signaling also requires that the efficiency of covert signaling, *r*, is sufficiently strong that at least *some* of the benefits of homophily can be capitalized upon, and this minimum efficiency increases with strength of homophily (Figure 2B).

It is worth noting that this last result concerning the efficiency of covert signaling, *r*, appears to contrast with analytical results from a similar model by Smaldino et al. (2018), who found that covert appeared to be *less* strongly favored under more efficient covert signals. However, that previous result now appears to be an artifact of the invasion criteria used in that model, whereby most agents

Figure 2

The Proportion of Agents Using Covert Signaling (A, C) and Churlish Receiving (B, D) at t = T Across 100 Simulations for Each Set of Parameter Values



Note. Covert signaling evolves when disliking costs are high and homophily is low (A), and when covert signaling is sufficiently efficient (C). Model parameters have a similar but less pronounced effect on receiving strategy, indicating that generous receiving coevolves with covert signaling (B, D). See the online article for the color version of this figure.

were assumed to be churlish by default and effects of drift were ignored. In our analysis of invasion and stability criteria (see Appendix), we find covert signaling reliably invades and can resist invasion by overt signaling when homophily is low, the disliking penalty is high, and the efficiency of covert signals is high.

Covert Signaling Coevolves With Generous Receiving

Figure 2 indicates that generous receiving coevolves with covert signaling, and churlish receiving coevolves with overt signaling. This result is made clearer in Figure 3, which correlates the levels of covert signaling and churlish receiving in each parameter condition from Figure 2. Covert signaling is associated with very high levels of generous receiving, while overt signaling is associated with moderately high levels of churlish receiving. This makes sense. If all receivers are churlish, covert signaling is never favored because there is no benefit to preventing dissimilar individuals from identifying that dissimilarity. Thus, in populations where covert signals are common, we should expect receiving to be generous. In contrast, when overt signaling is common, there is little benefit to being generous, especially if homophily is reasonably strong. Churlish receiving maximizes the likelihood that homophilic assortment will pair similar individuals and lowers the risk of assorting with a dissimilar partner due to lack of information.

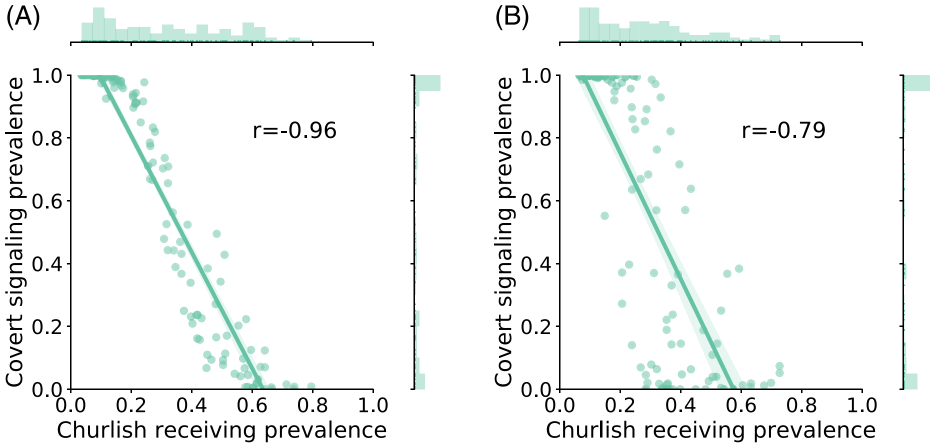
Covert Signaling Is Favored in More Diverse Populations

The similarity threshold, *S*, is the proportion of traits two individuals must have in common to be considered similar in our model. Because traits are randomly distributed in our model, one way to interpret this parameter is as a measure of the effective diversity of a population. When *S* is low, most pairs will be similar, even with random assortment. This could represent a population in which most differences are fairly trivial, and in which important traits regarding beliefs, norms, and goals tend to be shared. Larger values of *S* can represent more diverse populations. In such populations, the act of considering another individual to be similar requires sharing a large number of traits that represent suites of beliefs, norms, and goals.

We find that the evolution of covert signaling is quite responsive to this measure of diversity, moderated by the strength of homophily, *w* (Figure 4). When homophily is weak, the relationship between covert signaling and *S* is monotonically positive. That is, when homophily is weak—and people therefore often must interact with nonpreferred partners—covert signaling evolves more readily in more diverse populations. When most people are similar, selection on covert signaling is weaker because there is less of a need to avoid being disliked by dissimilar individuals; there are fewer of them.

Figure 3

Covert Signaling Thrives When There Is Little Churlish Receiving and Is Less Prevalent When Churlish Receiving Prevalence Increases

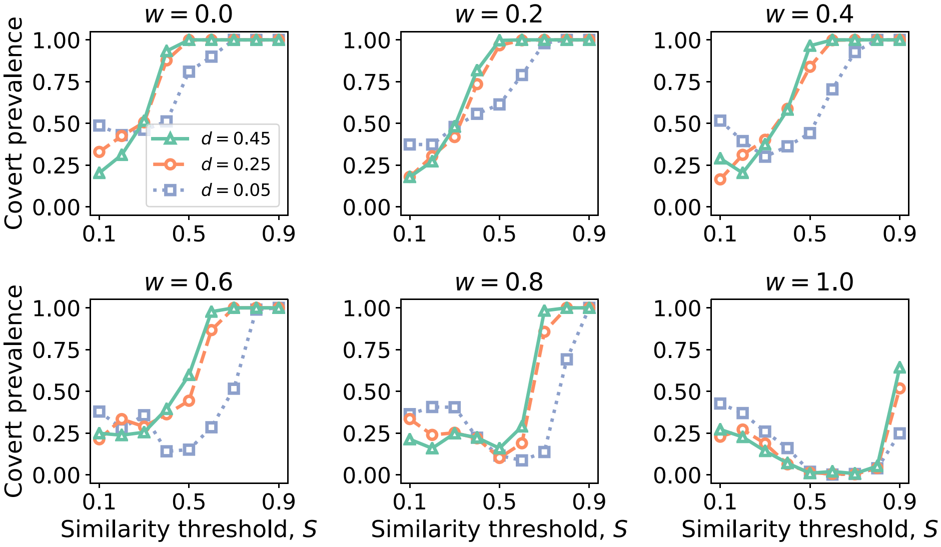


Note. The data are the values for covert signaling and churlish receiving in the paired heatmaps in figure 2, with (A) representing data in figure 2A, C and (B) representing data in figure 2 B, D. In each case, the line is the best-fit regression line, with the shaded region representing the 95% confidence interval. See the online article for the color version of this figure.

When homophily is strong, the relationship between covert signaling and the similarity threshold, *S*, becomes nonmonotonic. This is because when most potential pairs are similar, there is only very weak selection on signaling strategy, since most interactions will be between similar individuals. As *S* increases to moderate values, there is selection *against* covert signaling because strong homophily favors overt signaling when there is uncertainty about similarity. When *S* is very large, most potential partners will be dissimilar, and therefore avoiding occasional interactions with dissimilar partners is unavoidable. In this case, we once again get selection for covert signaling.

Figure 4

Homophily and Similarity



Note. When homophily is low to moderate (0.0 ≤ *w* ≤ 0.6), covert signaling prevalence increases monotonically with similarity threshold, *S*. For larger values of *w*, the prevalence of covert signaling decreases at first as a function of *S*, but then undergoes a transition, increasing toward a prevalence of 1 with all agents signaling covertly this monotonicity emerges earlier for lower values of *d*, the cost of disliking, for which selection on covert signaling is weaker. See the online article for the color version of this figure.

Covert Signaling Is Favored Among Minority Groups

The results showing that covert signals are favored in more diverse groups led us to hypothesize that covert signals would be preferentially favored among minority groups within a population—though, as noted earlier, this applies only to members of “invisible” minorities, whose minority status is not immediately apparent. In the absence of very strong homophily, minority individuals are more likely to find themselves partnered with dissimilar individuals, and therefore should have stronger incentives to conceal their full identities.

In our baseline model, all traits were uniformly distributed, and so all individuals had the same probability of being paired with a similar partner under random assortment. To test our hypothesis regarding minority individuals, we extended the model so that individuals were initialized as belonging to either the majority or the minority, where the minority constituted 10% of the population and the majority constituted the remaining 90%. These two groups differed on their first M traits, where $M = \frac{K-1}{2}$, so that members of the same group were highly likely to be similar to members of the same group and dissimilar to members of the other group, with the exact probabilities depending on both K and S. The first M traits of minority individuals were set to 1, while the same traits were set to 0 for members of the majority. Remaining traits were assigned at random as in the baseline model. So that individuals evolved strategies that were beneficial conditional upon their minority/majority status, targets for payoff-biased copying were restricted to agents with the same status as the focal agent.

The results of our simulations are depicted in Figure 5. We confirm that covert signaling is indeed more common among the minority under most conditions. For a high similarity threshold, S, covert signaling is so strongly favored under weak homophily that there is little difference between minority and majority individuals because all individuals have a sufficiently high probability of interacting with a dissimilar partner. For a moderate similarity threshold, covert signaling is always more prevalent among the minority than among the majority, providing strong support for our hypothesis.

Something interesting happens with a low similarity threshold ($S = 0.3$). The effect of more covert signaling among the minority is robust for low-to-moderate levels of homophily. However, the effect reverses when homophily is very strong, and covert signaling becomes relatively rare among the minority. This effect is driven by stronger selection for overt signaling on the minority—they need all the help they can get to find similar partners. When homophily is strong, both positive and negative attitudes are very useful, as they allow individuals both to find similar partners and to avoid dissimilar partners. Covert signals are problematic on both counts, as they reach fewer similar individuals and are completely missed by dissimilar individuals. This is true for majority individuals as well and is illustrated by the fact that covert signaling decreases for both minority and majority individuals in all cases as homophily increases. When the similarity threshold is low, however, majority individuals are likely to be similar to almost anyone they meet, and thus the selection for increased overt signaling can be stronger among members of the minority. Although we only considered one minority group here, our previous analyses—showing that covert signaling increases when diversity is high and all individuals are effectively members of a minority—indicate that this result should also hold when multiple minority groups coexist.

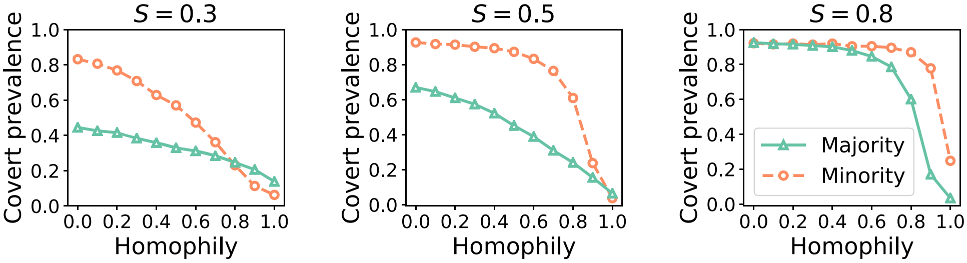
Our model assumes that the costs of disliking are equal for both majority and minority individuals. In reality, costs can differ between groups and are likely to be higher for minority individuals (Bunce & McElreath, 2018). Our earlier analyses of costs indicate that the effects of minority status, diversity, and differential cost should all be additive. That is, under higher costs, we should expect even more covert signaling among minority individuals, relative to majority individuals.

Discussion

Covert signaling allows individuals to assort with similar individuals when possible, while avoiding the potentially costs of detection by dissimilar individuals when assorting with them is unavoidable. The empirical examples we mentioned in our introduction—including dog whistles, humor, fashion, and activism—are among the best-known examples of covert signaling, but they have not, in general, been studied in a population context that would allow us to test the hypotheses generated by the model. Our model’s emphasis on the selective forces imposed by sociocultural and network factors provides new hypotheses to be tested. Specifically, we expect to see higher levels of covert signaling (relative to overt signaling) when the cost of being discovered as dissimilar is high, and when people are more likely to encounter dissimilar individuals in cooperative interactions.

Figure 5

Line Plots of Covert Signaling Prevalence in Minority and Majority Groups Over Homophily, for Three Values of Similarity Threshold, $S \in \{0.3,0.5,0.8\}$



Note. The minority group is 90% of the population. Disliking penalties are set to $d = \delta = 0.5$. See the online article for the color version of this figure.

The latter is especially likely to occur when assortment on attitudes is imperfect, when populations are diverse, and among invisible minorities. Scenarios that select for covert signaling also select for generous receiving—that is, being open minded toward strangers.

An implicit assumption of our approach is that much communication is about identity signaling. This is a view held by many others before us (e.g., Berger & Heath, 2008; Donath, 1999; Goffman, 1978), and one that we think is defensible. We make no normative statements about how people *should* signal their identity, overtly or covertly. What we *can* say, however, is that covert signals are adaptive in diverse populations where widespread cooperation is desirable and vehement disagreement is undesirable. Such populations were also associated with more generous attitudes toward unknown individuals in our model.

In the U.S., there is some evidence that generous attitudes toward unknown individuals are more common among political liberals, while conservatives may be more wary of strangers (Hibbing et al., 2014; Holbrook et al., 2017; Vigil, 2010). This seems consistent with the demographic differences between the parties—the left-leaning, urban-oriented Democratic party involves coalitions among disparate identity groups, while the increasingly far-right Republican party rallies around a more singular, exclusionary identity (Klein, 2020). Considering how strong political identity has become as a coalitional signal in the U.S. (Mason, 2018; Abramowitz & McCoy, 2019; Iyengar et al., 2019), individuals may only be incentivized to signal covertly to the extent they believe that cross-partisan cooperation is important and feasible.

Covert signaling and generous receiving involve being open to interactions with new, unknown people and an attitude of tolerance toward people who might be different (by masking those differences in mixed company). Covert signaling may therefore be part of what enables a healthy diverse, cosmopolitan society. In this respect, covert signaling bears a similarity to norms of politeness in that it may enable interactions between people who disagree on important issues by proscribing some overt utterances that would impede coordination or even foment conflict (Clark & Schunk, 1980; Yoon et al., 2020). It also should be stressed that generous receiving as discussed here does not imply higher-than-average tolerance of those who are different, only that penalties imposed on those who are different will be withheld until evidence for those differences is apparent. Thus, covert signaling may certainly be found under conditions in which receivers are generous in this way but where signalers nevertheless fear detection. Consider fashion choices used by oppressed minorities to find one another, or coded signals used by Cold War moles in order to out themselves to fellow spies. These individuals have reason to fear persecution should they be outed, and the fact that they can avoid persecution by signaling covertly indicates that their receivers are generous as we have defined it. Our formulation of churlishness in our model may also represent a somewhat extreme and even rare category of response, though one that may be more common in historically earlier or less globalized populations (Smaldino, 2019).

If covert signaling is favored when costs of being disliked are high and the ability to assort on identity is weak, what happens when these factors change? It is possible that as diverse societies become more segregated, and coalition-building with like-minded individuals becomes easier (as with social media and other internet-based mechanisms), selection for covert signaling may become weaker, and overt signals of identity more common. Indeed, changing conditions allowing for more efficient assortment and reduced costs for expressing dissimilarity may have driven the apparently sharp increase in racist hate speech that began to appear in the 2010s (Bjork-James & Maskovsky, 2017; Holt et al., 2015; Neiwert, 2018; Tucker et al., 2018; Youngblood, 2020). The rise of social media communities and the prominence of certain rallying public figures supporting racist rhetoric may have increased the opportunities for assortment while decreasing the perceived costs of being vilified by those who disagreed. The speed with which such communities arose suggests that, rather than people being converted to a cause, such communities were being maintained covertly and then began to feel emboldened to signal more overtly. Thus we see the rapid rise of outspoken radical groups and an increasingly polarized world. A world in which overt signaling is increasingly favored goes along with more churlish receiving. A world like this is either more segregated, more polarized, or both.

The widespread use of online platforms such as Facebook, Instagram, and Twitter—which serve as the infrastructure for much communication in the industrialized world—may favor conditions in which overt signaling is preferred. The ability to preferentially connect with like-minded individuals is a major incentive for using these platforms, and individuals can signal their suitability for belonging in specific communities by sharing information that confirms their allegiance to or suitability for those communities (Donath, 1999). Homophily in online networks is strong and the repercussions for being disliked by dissimilar individuals are often minimal. These conditions not only support polarization in the form of churlishness and hostility toward dissimilar individuals, but also incentivize misinformation as long as it reliably signals identity.

Covert signaling is a strategy that works best for identities that can be obscured. Obviously, not all identities are like this. Some, such as racial identities, are based on physical characteristics that are difficult or impossible to change. Nevertheless, covert signaling may occur among ethnically ambiguous individuals who wish to take advantage of multiple identities. Some groups, including some religious communities and criminal organizations, demand overt signals from their members, which serve explicitly to burn bridges with mainstream society[2] (Gambetta, 2009; Iannaccone, 1992; Sosis, 2003).

Our theory is agnostic about the precise nature of identity signals (e.g., words, images, fashions) and about the cognitive processes that produce and interpret them. This "behavioral gambit" is a standard technique in evolutionary and game-theoretic approaches to behavior (Grafen, 1991) that allows us to focus on the overarching strategies for signaling and interpreting identity information. We do hope that this work inspires closer investigation of the nature of covert and overt signals. From a cognitive perspective, current Bayesian approaches to pragmatic reasoning (e.g., Goodman & Frank, 2016) might be extended to include strategic behaviors on longer time scales to account for cultural variation in communication.

Our theory requires that covert signals remain obscured to dissimilar individuals. In cases where members of one group are incentivized to discover members of another group among them, members of the first group will eventually learn to identify

[2] On the other hand, covert signaling may be quite common among criminals, as their livelihood depends on not being detected as such (Gambetta, 2009).

members of the second group via their signaling behavior. In this way, initially covert signals may become increasingly overt. In a large and diverse population, at least some covert identity signals—likely those with cost asymmetries in being discovered—may exhibit cycles of chase-and-flight similar to those described by Simmel (1957) in his classic discussion of fashion cycles. As covert signals become less effectively covert, they may become repurposed as overt signals and while new covert signals emerge to replace them. The evolutionary dynamics of the signals used for both covert and overt signaling is an important topic for future study.

Our model design involves several simplifications that should be addressed. Our analysis focused on the use of honest signals of identity, and hence does not account for scenarios in which, for example, a member of a minority may overtly deceive listeners as to their true identity, or in which individuals may otherwise distort identity information to their advantage. Such behaviors are probably common and are likely to influence model dynamics. We ignored signaling errors other than those involving failure to perceive identity information, such as a failure to effectively conceal covert signals from dissimilar receivers. We also ignored payoff asymmetries that may be important for signaling strategies. For example, it is quite plausible that minority individuals may face more severe costs from being discovered as dissimilar than do members of a majority group. Future extensions of this model should investigate additional signaling and receiving strategies, errors, and payoff asymmetries, as the evolutionary fitness of a behavioral strategy is likely to be influenced by all of these factors (Boyd & Lorberbaum, 1987; McElreath & Boyd, 2007; Zefferman, 2014). Nevertheless, we believe that the model and analysis presented here provide an effective demonstration of the core theory of covert signaling.

The theory of covert signaling is inherently difficult to study empirically. This is partly because people often communicate differently depending on whom they are talking to as well as who else is listening (Bell, 1984; Clark & Carlson, 1982). More importantly, reliably detecting covert signaling by definition requires access to context typically possessed only by in-group individuals (Clark & Schaefer, 1987). This presents a problem for researchers not belonging to those in-groups. Nevertheless, some workarounds may exist. For example, one might consider well-established group identities and look for communications to which in-group members respond but out-group members ignore. It may also be possible to use field assistants from different identity groups, and compare differences in the identity information they observe from subjects. More generally, identifying covert signals is likely to involve a deep appreciation for pragmatics: Those aspects of communication in which meaning is highly dependent on context and shared knowledge. In anthropology, this is encompassed by the difference between etic and emic approaches—knowledge based on perspectives external and internal to a culture, respectively. Identifying covert signals may require at least some engagement with the emic perspective. For these reasons, covert identity signals may also be less discernible to machine-learning algorithms designed to infer identity. This implies that automated methods to detect social identity are likely to focus on those aspects which are signaled overtly. Such a scenario may be desirable for societal well-being.

References

Abramowitz, A., & McCoy, J. (2019). United States: Racial resentment, negative partisanship, and polarization in Trump's America. *The Annals of the American Academy of Political and Social Science*, *681*(1), 137–156. https://doi.org/10.1177/0002716218811309

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, *115*(3), 715–753. https://doi.org/10.1162/003355300554881

Aragones, E., & Neeman, Z. (2000). Strategic ambiguity in electoral competition. *Journal of Theoretical Politics*, *12*(2), 183–204. https://doi.org/10.1177/0951692800012002003

Barth, F. (1969). Introduction. In F. Barth (Ed.), *Ethnic groups and boundaries* (pp. 9–38). Little, Brown.

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145–204. https://doi.org/10.1017/S004740450001037X

Berger, J., & Heath, C. (2007). Where consumers diverge from others: Identity signaling and product domains. *Journal of Consumer Research*, *34*(2), 121–134. https://doi.org/10.1086/519142

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), Article 593. https://doi.org/10.1037/0022-3514.95.3.593

Berger, J. and Ward, M. (2010). Subtle signals of inconspicuous consumption. *Journal of Consumer Research*, *37*(4), 555–569. https://doi.org/10.1086/655445

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bjork-James, S., & Maskovsky, J. (2017). When white nationalism became popular. *Anthropology News*, *58*(3), e86–e91. https://doi.org/10.1111/AN.455

Bowles, S., & Gintis, H. (2013). *A cooperative species: Human reciprocity and its evolution*. Princeton University Press.

Boyd, R., & Lorberbaum, J. P. (1987). No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game. *Nature*, *327*(6117), 58–59. https://doi.org/10.1006/jtbi.1994.1092

Boyer, P. (2018). *Minds make societies: How cognition explains the world humans create*. Yale University Press.

Brooks, J. S., & Wilson, C. (2015). The inuence of contextual cues on the perceived status of consumption-reducing behavior. *Ecological Economics*, *117*, 108–117. https://doi.org/10.1016/j.ecolecon.2015.06.015

Bunce, J. A., & McElreath, R. (2018). Sustainability of minority culture when inter-ethnic interaction is profitable. *Nature Human Behaviour*, *2*(3), 205–212. https://doi.org/10.31235/osf.io/bpgt3

Calcott, B. (2008). The other cooperation problem: Generating benefit. *Biology & Philosophy*, *23*(2), 179–203. https://doi.org/10.1007/s10539-007-9095-5

Castro, L., & Toro, M. A. (2007). Mutual benefit cooperation and ethnic cultural diversity. *Theoretical Population Biology*, *71*(3), 392–399. https://doi.org/10.1016/j.tpb.2006.10.003

Chwe, M. S.-Y. (2001). *Rational ritual: Culture, coordination, and common knowledge*. Princeton University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In Lauren B. Resnick, John M. Levine, & Stephanie D. Teasley (Eds.), *Perspectives on socially shared cognition*. American Physiology Association.

Clark, H. H., & Carlson, T. B. (1982). Hearers and speech acts. *Language*, *58*, 332–373. https://doi.org/10.2307/414102

Clark, H. H., & Schaefer, E. F. (1987). Concealing one's meaning from overhearers. *Journal of Memory and Language*, *26*(2), 209–225. https://doi.org/10.1016/0749-596X(87)90124-0

Clark, H. H., & Schunk, D. H. (1980). Polite responses to polite requests. *Cognition*, *8*(2), 111–143. https://doi.org/10.1016/0010-0277(80)90009-8

Donath, J. S. (1999). Identity and deception in the virtual community. In P. Kollock, & M. Smith (Eds.), *Communities in cyberspace* (pp. 29–59). Routledge.

Efferson, C., Lalive, R., & Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism. *Science*, *321*(5897), 1844–1849. https://doi.org/10.1126/science.1155805

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. *Communication Monographs*, *51*(3), 227–242. https://doi.org/10.1080/03637758409390197

Farrell, J. and Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, *10*(3), 103–118. https://doi.org/10.1257/jep.10.3.103

Fischer, H. (2015). *Gay semiotics: A photographic study of visual coding among homosexual men*. Cherry and Martin.

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. https://doi.org/10.1556/JEP.6.2008.4.2

Flamson, T., & Barrett, H. C. (2013). Encrypted humor and social networks in rural Brazil. *Evolution and Human Behavior*, *34*(4), 305–313. https://doi.org/10.1016/j.evolhumbehav.2013.04.006

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* (Vol. 1, pp. 49–73). John Benjamins Publishing.

Gallotti, M., & Frith, C. D. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences*, *17*(4), 160–165. https://doi.org/10.1016/j.tics.2013.02.002

Gambetta, D. (2009). *Codes of the underworld: How criminals communicate*. Princeton University Press.

Goffman, E. (1978). *The presentation of self in everyday life*. Harmondsworth.

Gold, S. J. (1991). Ethnic boundaries and ethnic entrepreneurship: A photo-elicitation study. *Visual Studies*, *6*(2), 9–22. https://doi.org/10.1080/14725869108583688

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*(4), 517–546. https://doi.org/10.1016/S0022-5193(05)80088-8

Grafen, A. (1991). Modelling in behavioural ecology. In J. R. Krebs, & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (pp. 5–31). Wiley.

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics: Speech acts* (pp. 41–58). Brill.

Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D. L., Edmonds, B., Ge, J., Giske, J., Groeneveld, J., et al. (2020). The odd protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *Journal of Artificial Societies and Social Simulation*, *23*(2), 7. https://doi.org/10.18564/jasss.4259

Haney-López, I. (2015). *Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class*. Oxford University Press.

Henderson, R., & McCready, E. (2017). How dogwhistles work. In K. Mineshima, K. Kojima, K. Satoh, S. Arai, D. Bekki, & Y. Ohta (Eds.), *JSAI International Symposium on Artificial Intelligence* (pp. 231–240). Springer.

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. https://doi.org/10.1017/S0140525X13001192

Holbrook, C., López-Rodríguez, L., Fessler, D. M. T., Vázquez, A., & Gómez, A. (2017). Gulliver's politics: Conservatives envision potential enemies as readily van-quished and physically small. *Social Psychological and Personality Science*, *8*(6), 670–678. https://doi.org/10.1177/1948550616679238

Holt, T., Freilich, J. D., Chermak, S., & McCauley, C. (2015). Political radicalization on the internet: Extremist content, government control, and the power of victim and jihad videos. *Dynamics of Asymmetric Conict*, *8*(2), 107–120. https://doi.org/10.1080/17467586.2015.1065101

Hutchison, J., & Martin, D. (2015). The evolution of stereotypes. In V. Zeigler-Hill, L. L. M. Welling, & T. K. Shackelford (Eds.), *Evolutionary perspectives on social psychology* (pp. 291–301). Springer.

Iannaccone, L. R. (1992). Sacrifice and stigma: Reducing free-riding in cults, communes, and other collectives. *Journal of Political Economy*, *100*(2), 271–291. https://doi.org/10.1086/261818

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*, 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Klein, E. (2020). *Why we're polarized*. Simon & Schuster.

Kuran, T. (1989). Sparks and prairieffres: A theory of unanticipated political revolution. *Public Choice*, *61*(1), 41–74. https://doi.org/10.1007/BF00116762

Kuran, T. (1995). The inevitability of future revolutionary surprises. *American Journal of Sociology*, *100*(6), 1528–1551. https://doi.org/10.1086/230671

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*(3), 785. https://doi.org/10.1037/a0019688

Leong, C.-H., Komisarof, A., Dandy, J., Jasinskaja-Lahti, I., Safdar, S., Hanke, K., & Teng, E. (2020). What does it take to become one of us?" redefining ethnic-civic citizenship using markers of everyday nationhood. *International Journal of Intercultural Relations*, *78*, 10–19. https://doi.org/10.1016/j.ijintrel.2020.04.006

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, *54*(4), 421–431. https://doi.org/10.1007/s10539-006-9049-3

Loury, G. C. (1994). Self-censorship in public discourse: A theory of "political correct-Ness" and related phenomena. *Rationality and Society*, *6*(4), 428–461, https://doi.org/10.1177/1043463194006004002

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. University of Chicago Press.

McElreath, R. and Boyd, R. (2007). *Mathematical models of social evolution: A guide for the perplexed*. University of Chicago Press.

McElreath, R., Boyd, R., & Richerson, P. J. (2003). Shared norms and the evolution of ethnic markers. *Current Anthropology*, *44*(1), 122–130. https://doi.org/10.1086/345689

Miller, G. F., & Todd, P. M. (1998). Mate choice turns cognitive. *Trends in Cognitive Sciences*, *2*(5), 190–198. https://doi.org/10.1016/S1364-6613(98)01169-3

Moffett, M. W. (2019). *The human swarm: How our societies arise, thrive, and fall*. Basic Books.

Nee, V., Sanders, J. M., & Sernau, S. (1994). *Job transitions in an immigrant metropolis: Ethnic boundaries and the mixed economy* (pp. 849–872). American Sociological Review.

Neiwert, D. (2018). *Alt-America: The rise of the radical right in the age of Trump*. Verso Books.

Nettle, D., & Dunbar, R. (1997). Social markers and the evolution of reciprocal exchange. *Current Anthropology*, *38*, 93–99. https://doi.org/10.1086/204588

Owings, D. H., & Morton, E. S. (1997). The role of information in communication: An assessment/management approach. In D. H. Owings, M. D. Beecher, & N. S. Thompson (Eds.), *Communication* (pp. 359–390). Springer.

Sanders, J. M. (2002). Ethnic boundaries and identity in plural societies. *Annual Review of Sociology*, *28*(1), 327–357. https://doi.org/10.1146/annurev.soc.28.110601.140741

Searle, J. R. (1975). Indirect speech acts. In P. Cole, & J. L. Morgan (Eds.), *Speech acts* (pp. 59–82). Brill.

Shelp, S. G. (2003). Gaydar. *Journal of Homosexuality*, *44*(1), 1–14. https://doi.org/10.1300/J082v44n01_01.

Silk, J. B., Kaldor, E., & Boyd, R. (2000). Cheap talk when interests conict. *Animal Behaviour*, *59*(2), 423–432. https://doi.org/10.1006/anbe.1999.1312

Simmel, G. (1957). Fashion. *American Journal of Sociology*, *62*(6), 541–558. https://doi.org/10.1086/222102

Smaldino, P. E. (2014). The cultural evolution of emergent group-level traits. *Behavioral and Brain Sciences*, *37*, 243–295. https://doi.org/10.1017/S0140525X13001544

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Social Psychology* (pp. 311–331). Routledge.

Smaldino, P. E. (2019). Social identity and cooperation in cultural evolution. *Behavioural Processes*, *161*, 108–116. https://doi.org/10.1016/j.beproc.2017.11.015

Smaldino, P. E. (2020). How to translate a verbal theory into a formal model. *Social Psychology*, *51*(4), 207–218. https://doi.org/10.31222/osf.io/n7qsh

Smaldino, P. E., Flamson, T. J., & McElreath, R. (2018). The evolution of covert signaling. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-22926-1

Smaldino, P. E., Janssen, M. A., Hillis, V., & Bednar, J. (2017). Adoption as a social marker: Innovation diffusion with outgroup aversion. *Journal of Mathematical Sociology*, *41*(1), 26–45. https://doi.org/10.1080/0022250X.2016.1250083

Sosis, R. (2003). Why aren't we all hutterites? *Human Nature*, *14*(2), 91–127. https://doi.org/10.1007/s12110-003-1000-6

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*, 355–374. https://doi.org/10.2307/1882010

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Harvard University Press Cambridge.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691. https://doi.org/10.1017/S0140525X05000129

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). *Social media, political polarization, and political disinformation: A review of the scientific literature*. Hewlett Foundation. https://ssrn.com/abstract=3144139

Vigil, J. M. (2010). Political leanings vary with facial expression processing and psychosocial functioning. *Group Processes & Intergroup Relations*, *13*(5), 547–558. https://doi.org/10.1177/1368430209356930

Wimmer, A. (2013). *Ethnic boundary making: Institutions, power, networks*. Oxford University Press.

Yoon, E. J., Tessler, M. H., Goodman, N. D., and Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, *4*, 71–87. https://doi.org/10.1162/opmi_a_00035

Youngblood, M. (2020). Extremist ideology as a complex contagion: The spread of far-right radicalization in the united states between 2005-2017. *Humanities & Social Sciences Communication*, *7*, 49. https://doi.org/10.1057/s41599-020-00546-3

Zefferman, M. R. (2014). Direct reciprocity under uncertainty does not explain oneshot cooperation, but demonstrates the benefits of a norm psychology. *Evolution and Human Behavior*, *35*(5), 358–367. https://doi.org/10.1016/j.evolhumbehav.2014.04.003

Appendix A

Assortment Algorithm

In this section, we provide the details of the algorithm by which agents in our model assort for cooperative interactions.

For each of n_R rounds of cooperative assortment per time step, agents pair up as follows. The set of available agents, $\mathcal{A}$, contains those agents not yet paired. First, a focal agent is selected at random and removed from the set of available agents. Then, probabilities are calculated for interacting with all other available agents. When homophily (w) is zero, all available agents have equal probability of pairing. Otherwise, agents are more likely to pair with agents whom they like and who like them in return, and less likely to pair with agents whom they dislike and who dislike them in return. For each available agent j, the focal agent i calculates an *interaction factor* f_{ij}, defined as

$$f_{ij} = 1 + \frac{w(a_{ij} + a_{ji})}{2}, \tag{A1}$$

where w is the global homophily level parameter; a_{ij} and a_{ji} are i's attitude toward j and vice versa (see Table 1). The probability focal agent i interacts with potential partner j is

$$\Pr(i\,\text{pairs with}\,j) = \frac{f_{ij}}{\sum\limits_{j \in \mathcal{A}} f_{ij}}. \tag{A2}$$

Agent i's partner is drawn according to these probabilities. Both agents are removed from the set of available agents and the process continues with a new focal agent until all agents are paired.

Appendix B

Supplemental Analyses

Here, we explore the temporal dynamics of the model and the conditions for invasion and evolutionary stabilities when particular signaling or receiving strategies are initially rare. We also ran a number of additional analyses to assess the sensitivity and robust of the model, and we present the results of these analyses here. We generally conclude that the model is quite robust.

Invasion and Evolutionary Stability

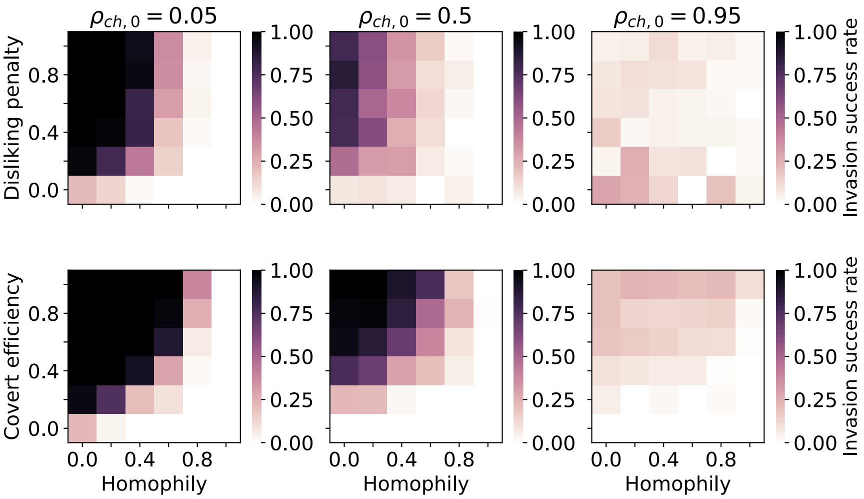
The simulations reported in the main text were initialized with both covert and overt signaling strategies used by 50% of the population. Here, we check the robustness of these results by considering populations in which one signaling strategy is initially dominant (at 95%) of the population, and we consider the spread of initially rare users of the other signaling strategy (5% of the population). An invading strategy was determined to have successfully invaded if, after $T = 100$ time steps, it had increased in frequency from 5%. Otherwise, the dominant strategy resisted invasion.

Figure B1 shows the conditions for the invasion of covert signaling under default conditions, as a function of homophily (w), disliking penalty ($d = \delta$), and the efficient of covert signaling (r/R). The first column shows that when most receivers are generous (i.e., the initial frequency of churlish receivers, $\rho_{ch,0}$, is low), covert signaling readily invades as long as homophily is not too strong, the

(*Appendices continue*)

Figure B1

Heatmaps Show the Success Rate for Covert Signaling Invasion for Different Initial Prevalences of Churlish Receiving, Denoted $\rho_{ch,0}$ *in the Heatmap Column Titles, for Several Settings of Disliking Penalty d or Covert Signaling Efficiency, R, and Homophily w*

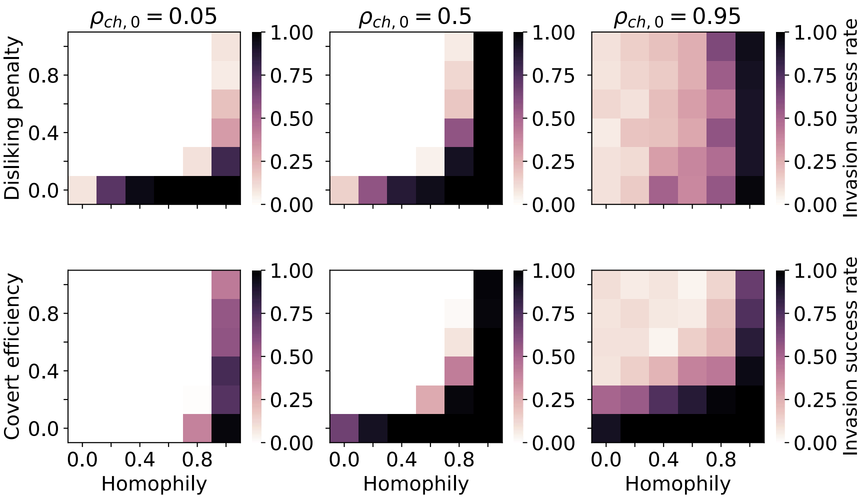


Note. In these experiments, the initial prevalence of covert signalers is set to 0.05. If the prevalence rises to an equilibrium above 0.05, invasion is considered successful we see again that generous receiving favors covert signaling, but higher initial churlish receiving prevalence takes away the adaptive benefits of covert signaling. See the online article for the color version of this figure.

dislike penalty is sufficiently high, and covert signals are sufficiently efficient. The middle and right columns show that when more receivers are initially churlish, covert signaling is less likely to invade. Figure B2 shows that complementary results hold for the invasion of overt signaling when covert signaling dominates. Overt signaling can invade when homophily is strong, disliking penalties are small, covert signals are inefficient, or most receivers are churlish.

Figure B2

Heatmaps Show the Success Rate for Overt Signaling Invasion for Different Initial Prevalences of Churlish Receiving, Denoted $\rho_{ch,0}$ *in the Heatmap Column Titles, for Several Settings of Disliking Penalty d or Covert Signaling Efficiency, R, and Homophily w*



Note. In these experiments, the initial prevalence of covert signalers is set to 0.05. If the prevalence rises to an equilibrium above 0.05, invasion is considered successful we see again that generous receiving favors covert signaling, but higher initial churlish receiving prevalence takes away the adaptive benefits of covert signaling. See the online article for the color version of this figure.

(*Appendices continue*)

Varying the Costs of Being Disliked, *d* and δ

In the main analysis, we assumed that the disliking penalty, *d*, and synergistic cost of mutual dislike, δ were equal, such that the cost of mutual dislike was always twice the cost of one person disliking the other. Here, we test different assumptions about the nonequality of *d* and δ. First, we set δ to be one of a small set of values and rerun our analyses letting *d* vary as *d* = δ varied before (Figure B3). As expected, we see more covert signaling as a function of increased δ, but *not* as a pure function of *d* when δ is held constant. In other words, *it is the cost of mutual dislike that primarily drives the evolution of covert signaling*. We confirm this by holding *d* constant and varying δ (Figure B4).

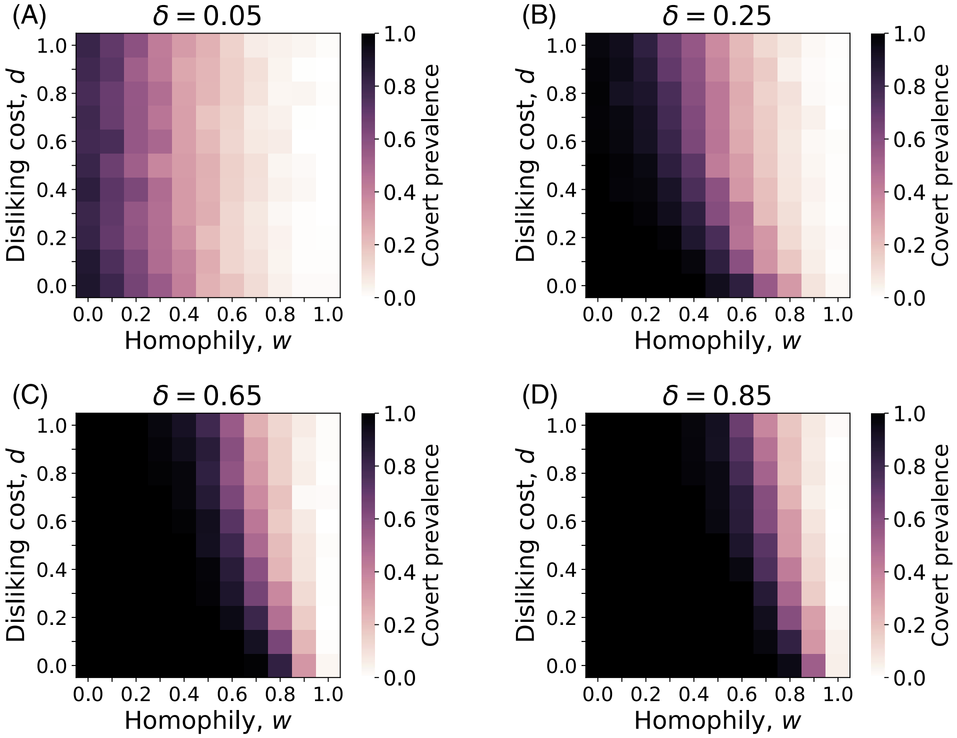
Why does this occur? Consider the case of constant δ and varying *d*. By sending encrypted signals, a covert signaler may avoid being disliked, but this actually *increases* their chance of interacting with someone who is dissimilar when homophily is imperfect. In other words, covert signalers are more likely to interact with someone who is different. If paired with an overt signaler, the covert signaler is more likely to dislike their partner. In this case, if δ is small, the covert signaler is not incentivized to avoid being disliked (because they are already disliked), but they *are* incentivized to become better at avoiding those they dislike or who might dislike them. In other words, to signal overtly. This means that the selection pressures of *d* and δ actually work in opposite directions. We believe that our assumption that mutual dislike imposes a higher cost than unilateral dislike is a reasonable one, but it is worth noting that our model's predictions regarding the cost of being disliked depend on this assumption.

Varying the Similarity Bonus, *S*

In the main text analyses, we used a similarity bonus of *s* = 0.25, meaning that being paired with a similar partner provided a payoff 25% higher than interacting with a dissimilar partner. We find that our results are robust to other values of the benefit to similarity. Unsurprisingly, greater benefits to similarity incentivize mechanisms that increased the chance assortment, favoring overt signaling. Decreased benefits to similarity strengthen selection on mechanisms that allow individuals to avoid being disliked, favoring covert signaling (Figure B5).

Figure B3

Heatmaps Show Prevalence of Covert Signaling for Various Parameter Settings of d, δ*, and w*

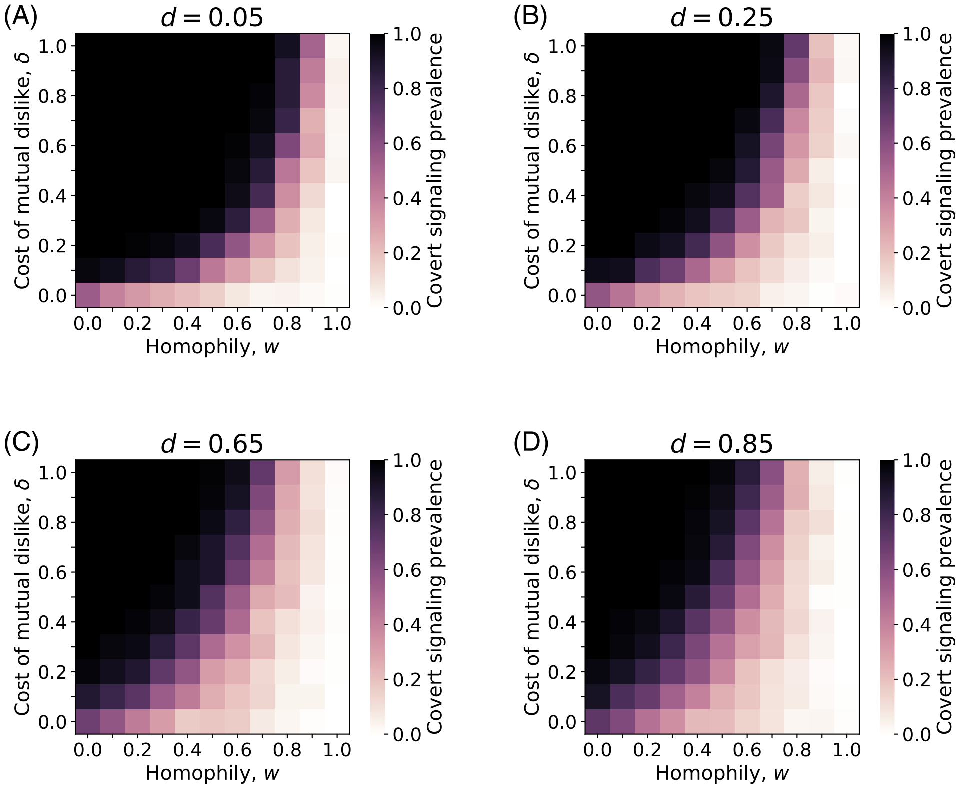


Note. To generate the results shown here we set the cost of mutual dislike δ to be constant (shown in subfigure title) and varied *d*, as shown on the *y*-axis of the subfigures as δ increases, covert signaling prevalence increases overall across parameter settings. See the online article for the color version of this figure.

(*Appendices continue*)

Figure B4

Heatmaps Show Prevalence of Covert Signaling for Various Parameter Settings of d, δ, and w



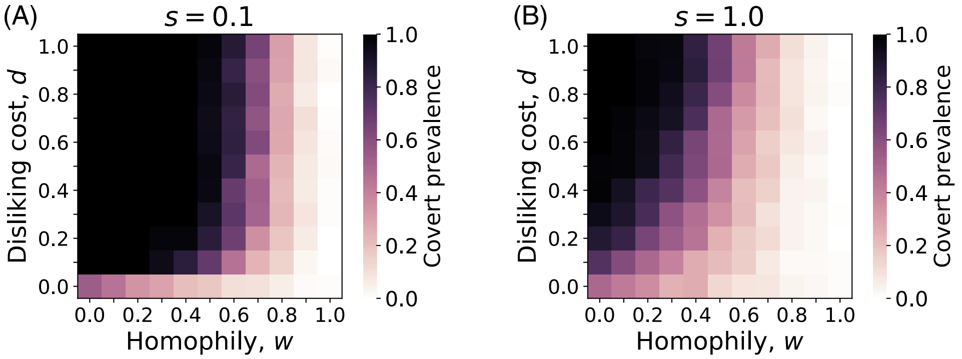
Note. To generate the results shown here we set the disliking cost *d* to be constant (shown in subfigure title) and varied δ, as shown on the *y*-axis of the subfigures as *d* increases, covert signaling prevalence decreases overall across the other parameter settings. See the online article for the color version of this figure.

Varying the Efficiency of Overt Signals, *R*

In the main text analyses, we assumed perfect transmission for overt signals, $R = 1$. In other words, all overt signals were perfectly transmitted to all receivers, and covert signals were transmitted to a smaller fraction of (only similar) receivers, *r*. In Figure B6 we show the results for two other values of *R*, holding *R* fixed at *R*/2. This analysis demonstrates that our results are quite robust to smaller values of *R* and that it appears to be the *relative* efficiency of covert signals, *r*/*R*, that matters, as shown in the main text.

Figure B5

Heatmaps Show Average Prevalence of Covert Signaling for Two Settings of the Similarity Bonus, s

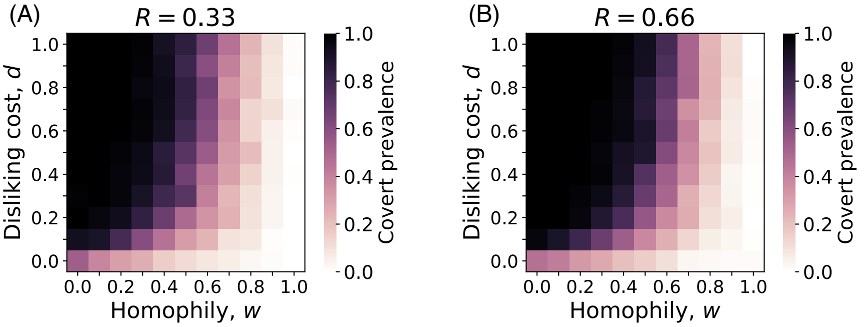


Note. All other parameters were set to their default values if not otherwise indicated. See the online article for the color version of this figure.

(*Appendices continue*)

Figure B6

Heatmaps Show Average Prevalence of Covert Signaling for Two Settings of the Efficiency of Overt Signals, R



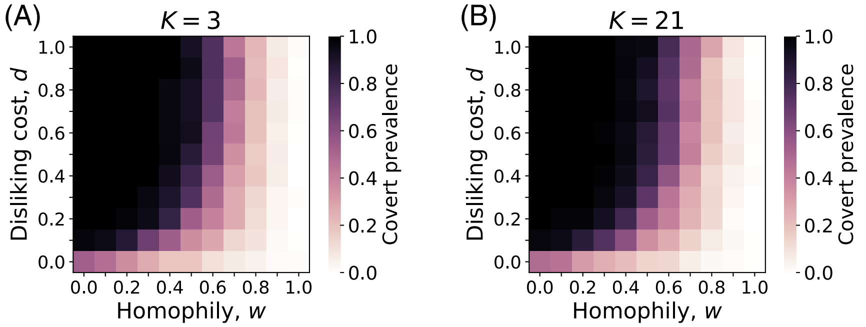
Note. r was rescaled so that the value of the ratio between r and R is the same as in the main text when $R = 1$, $r/R = 0.5$. All other parameters were set to their default values if not otherwise indicated. See the online article for the color version of this figure.

Varying the Number of Traits, *K*

In the main text analysis, we showed only results for $K = 9$ traits. We reproduced many of our analyses for a wide range of values of K. Our results are largely insensitive to the exact number of traits, as shown by the replications of our main result for $K = \{3, 21\}$ in Figure B7. Results that were dependent on the similarity threshold, S, were slightly affected by K because it determined the precise number of traits needed to reach specific thresholds. The qualitative patterns still held for all values of K, though. Figure B8 replicates part of the main text Figure 4 for different values of K showing this qualitative robustness.

Figure B7

Heatmaps Show Average Prevalence of Covert Signaling for Two Settings of the Number of Agent Traits, K

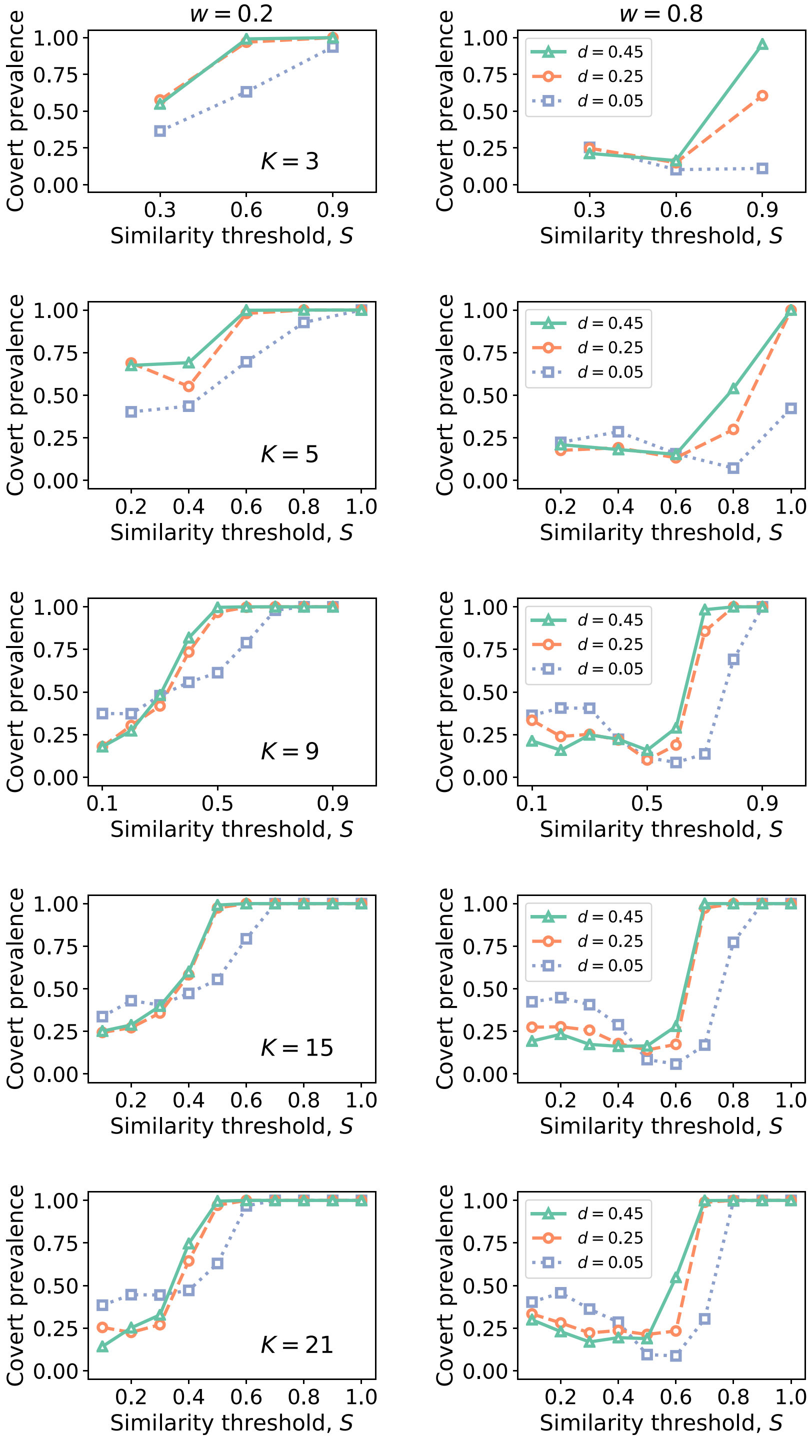


Note. All other parameters were set to their default values if not otherwise indicated. See the online article for the color version of this figure.

(*Appendices continue*)

Figure B8

Reproduction of Main Text Figure 4 for Different Values of K, for High and Low Values of Homophily Only (w = {0.2,0.8})



Note. See the online article for the color version of this figure.

Received December 11, 2020
Revision received October 7, 2021
Accepted October 10, 2021 ■