Hipsters and the Cool: A Game Theoretic Analysis of Social Identity, Trends and Fads

Russell Golman*[1], Erin H. Bugbee[1], Aditi Jain[2], and Sonica Saraf[3]

[1]Department of Social and Decision Sciences, Carnegie Mellon University
[2]Department of Mathematics, Carnegie Mellon University
[3]Center for Neural Science, New York University

December 22, 2020

Abstract

Cultural trends and popularity cycles can be observed all around us, yet our theories of social influence and identity expression do not explain what perpetuates these complex, often unpredictable social dynamics. We propose a theory of social identity expression based on the opposing, but not mutually exclusive, motives to conform and to be unique among one's neighbors in a social network. We then model the social dynamics that arise from these motives. We find that the dynamics typically enter random walks or stochastic limit cycles rather than converging to a static equilibrium. The dynamics also exhibit momentum, preserve diversity, and usually produce more conformity between neighbors, in line with empirical stylized facts. We also prove that without social network structure or, alternatively, without the uniqueness motive, reasonable adaptive dynamics would necessarily converge to equilibrium. Thus, we show that nuanced psychological assumptions (recognizing preferences for uniqueness along with conformity) and realistic social network structure are both critical to our account of the emergence of complex, unpredictable cultural trends.

Keywords: Conformity | Games on Social Networks | Popularity Cycles | Social Dynamics | Uniqueness

*Corresponding Author; E-mail: rgolman@andrew.cmu.edu

Introduction

Popular cultural practices come into and out of fashion. Researchers have observed boom-and-bust cycles of popularity in music, clothing styles, automobile designs, home furnishings, given names, and even management practices (Shuker, 2016; Richardson and Kroeber, 1940; Reynolds, 1968; Sproles, 1981; Berger, 2008; Berger and Le Mens, 2009; Lieberson, 2000; Lieberson and Lynn, 2003; Abrahamson, 1991; Zuckerman, 2012). Popularity cycles appear to be driven by social influence, e.g., by people adopting the music that their friends listen to or that they perceive as popular (Salganik et al., 2006; Salganik and Watts, 2008). At the individual level, people are constantly looking for new ways to express their preferred social identities (Hetherington, 1998; Rentfrow and Gosling, 2006; Berger, 2008; Chan et al., 2012). The resultant social dynamics do not typically converge to equilibrium. What are the social forces that lead to such perpetual change and novelty?

Social pressure to conform is a powerful force when behavioral patterns across a society shift in unison. Psychologists since Asch have recognized the remarkable strength of the conformity motive, stemming from a fundamental goal to fit in as part of a social group (Asch, 1955, 1956; Cialdini and Trost, 1998). People tend to feel uncomfortable about considering, holding, and expressing beliefs that conflict with the prevailing views around them as well as about behaving oddly, in ways that might expose oneself as an outsider to the group (Turner et al., 1987; Golman et al., 2016). Given the conformity motive alone, we might expect to observe convergence to an equilibrium in which society becomes monolithic, yet instead we actually observe persistent diversity.

Opposing the motive to conform is a similarly universal human need for uniqueness (Snyder and Fromkin, 1980; Lynn and Snyder, 2002). While the desire to differentiate oneself clearly works against the desire to blend in (Imhoff and Erb, 2009), Chan, Berger and van Boven (2012) demonstrate that people simultaneously pursue assimilation and differentiation goals, aiming to be identifiable, but not identical (see also Leibenstein, 1950). Preferences for idiosyncratic behavioral patterns can preserve diversity (Smaldino and Epstein, 2015). Still, the question remains why behavioral patterns often do not remain in a stable equilibrium with everyone finding an optimal balance between distinctiveness and conformity. Why instead do behavioral patterns go through perpetual change, with particular behaviors cycling into and out of fashion as cultural trends play out?

One explanation, tracing back to Simmel (1957), is that an upper class tries to distinguish itself from the common folk while the common folk try to imitate them. In modern models of identity signaling, membership in one group may be preferable to membership in another, and people want to strategically distinguish themselves from those in the less favorable group (Berger and Heath, 2007). The resulting dynamic of imitation and differentiation (or "chase-and-flight") can lead to fashion cycles (Pesendorfer, 1995; Bakshi et al., 2013). (Relatedly, games like "matching pennies" also generate best-response cycles involving imitation and differentiation, which have been associated with fashion cycles (Karni and Schmeidler, 1990; Zhang et al., 2018), but the confined strategy space in these games leaves little room for the kind of unpredictable boom-and-bust cycles we explore here.) Undoubtedly, there are contexts in which elites initiate fashions and everyone else strives to imitate them, but empirical research shows that in many other contexts, groups with lower or equal status also strive to differentiate themselves (Berger and Heath, 2008). A dynamic of mutual differentiation, without imitation, cannot account for popularity cycles.

Other models of popularity cycles rely on people continually discovering new behaviors, which spread through the population and then get discarded, either through random imitation (Bentley et al., 2004, 2007), or with a motive for conformity or anti-conformity (Acerbi and Bentley, 2014), or with the co-evolution of behavior and preferences (Acerbi et al., 2012). These models account for boom-and-bust cycles of popularity, but do not attempt to explain the source of the new behaviors that continually enter the model and keep the dynamics from converging to equilibrium.

This paper explores a new account of the dynamics of cultural trends and popularity cycles. We show that along with conformity and uniqueness motives, a realistic network of social interaction may be a critical ingredient for complex social dynamics to emerge. Specifically, we show that reasonable adaptive dynamics that would necessarily converge to a static equilibrium given random interactions in a well-mixed pool of people instead typically enter random walks or stochastic limit cycles, and thus never converge, when interactions are restricted to individuals' local neighborhoods in their social networks.

Popularity cycles in the expression of social identities display a number of empirical regularities, beyond the simple observation that they do not converge to equilibrium. They often preserve diversity, with different people expressing different identities. A hallmark pattern of social influence, that friends or acquaintances tend to behave similarly, holds for identity expression as well as other kinds of behaviors (Christakis and Fowler, 2013), and commonalities can extend across large communities. For example, there are regional correlations in the frequencies of given names across U.S. states (Barucca et al., 2015). Non-controversial behaviors often spread most quickly through "weak ties" in loosely clustered networks (the strength of the weak tie being its tendency to serve as a bridge between groups with otherwise limited contact), whereas behaviors that require social reinforcement from multiple sources, e.g., innovative health behaviors or participation in social movements, tend to spread more quickly through more tightly clustered social networks, in a process of "complex contagion" (Centola and Macy, 2007; Centola, 2010). As contagions spread, popularity cycles exhibit momentum – changes in popularity tend to persist in the same direction over time (Gureckis and Goldstone, 2009). Moreover, consistent with the motives we assume for our model, trends of rising popularity may spill over to other similar, but not identical, expressions of identity, while over-popularity actually decreases further adoption of particular expressions of identity (Berger et al., 2012). Here we find that the social dynamics that emerge in our model with social network structure exhibit momentum, preserve within-group diversity, and usually produce more conformity between network neighbors.[1]

A natural theoretical approach for investigating social influence on decisions is to use game theory. The conformity motive in isolation would create a Keynesian beauty contest, in which what is cool (like what is beautiful) is just what everybody else believes is cool (Keynes, 1936). The

[1] In contrast, chase-and-flight dynamics between stratified social classes do not preserve diversity within the class that is trying to imitate the elite. And models that assume completely random drift cannot account for the empirical pattern that popularity cycles exhibit momentum.

uniqueness motive in isolation would create a congestion game, in which the objective is simply to be distinct from as many other people as possible (Rosenthal, 1973). Both games are known to be potential games, for which convergence to a pure strategy Nash equilibrium is practically guaranteed (Monderer and Shapley, 1996a,b). When both motives co-exist and the game is played on a realistic social network, however, the dynamics are more complex.

Cultural trends can be modeled more realistically as the dynamics of a game on a social network because social influence is mediated by a social network. Social influence on expressions of individual identity is transmitted whenever an individual observes another person whom he would like to identify with, so the relevant social network is defined by directed connections corresponding to observation. The connected components of the social network may correspond to distinct social groups, each with its own emergent subculture.

The desire for uniqueness within one's own social group should not be conflated with a desire for differentiation across groups (Chan et al., 2012). Our model features in-group conformity and uniqueness motives; it could be augmented with a desire for differentiation across groups, but for parsimony we assume that people care only about their fit within their own groups.

Model 1: Social Identity Expression in a Well-Mixed Population

We model the expression of social identity as a game played by a population of N individuals. Let us say there are m aspects of identity (or identity-relevant traits). Each person i adopts an expression of identity $x_i = x_{i,1}, ..., x_{i,m}$, where the choice of each expressed trait $x_{i,\mu} \in \{a..b\}^d$ can be represented as a tuple of d integers from some interval.[2] For example, in the case of choosing an outfit to wear, two traits could be the color of the shirt and the color of the pants, and three integers between 0 and 255 might correspond to shades of red, green, and blue that mix together to form any color in an RGB color system.

A person's degree of conformity in the population depends on the (Euclidean) distance between

[2] The dimensionality d of the tuple and the boundaries of the interval $a..b$ can certainly vary for different traits, but we omit subscripts on these parameters specifying a particular trait to simplify the notation.

his expressed identity and the average (population mean) expression of identity, $\|x_i - \bar{x}\|$. A person's degree of uniqueness in the population depends on the number of others who express the exact same identity-relevant trait as him, averaged across all traits. For individual i and trait μ, denote the number of others who adopt his exact same expression of this trait as $n_{i,\mu}(X)$, where X is the entire population's profile of expressed identities, and let $n_i(X)$ denote the average amount of shared traits (i.e., $n_i(X) = \frac{1}{m}\sum_\mu n_{i,\mu}(X)$). Putting together the conformity and uniqueness motives, we model person i's utility given the profile of expressed identities as

$$u_i(X) = -\|x_i - \bar{x}\|^2 - \lambda\, n_i(X) \tag{1}$$

where λ is a parameter that describes the strength of the uniqueness motive relative to the conformity motive. This utility function describes a person whose goal is to be similar to everybody, yet the same as nobody (Chan et al., 2012).

Over time people may change their expressions of identity to achieve higher utility. We need not fully prescribe this process, but assume only that people make changes that increase their own utility, in accordance with some *better-reply dynamics* (Monderer and Shapley, 1996b; Friedman and Mezzetti, 2001).

Definition 1 (Better-reply dynamics)**.** *At any given time t, one person i may consider switching from x_i to x_i'; he switches if and only if $u_i(X') > u_i(X)$; and for each person i and any best response x_i^* (to $X(t)$), the expected time until person i considers switching to x_i^* is finite.*

The motivation for better-reply dynamics is that people are boundedly rational and adaptive (Gigerenzer, 2000). They can see what the people around them are doing and can search for something better (myopically), but they do not instantaneously react to changes in other people's behavior or anticipate these changes before they occur (Fiske and Taylor, 2013). Almost all commonly assumed adaptive learning dynamics are particular specifications of better-reply dynamics (Hofbauer and Sigmund, 2003).

Results: Social Dynamics in a Well-Mixed Population

Theorem 1. *Suppose people derive utility from both their conformity and their uniqueness in the population, as in Equation (1). Then any better-reply dynamics necessarily converge to a pure strategy Nash equilibrium.*

The proof is presented in the SM Appendix. It follows from Lemma 1 in the SM Appendix, which identifies an exact potential function for this game. Two examples of Nash equilibria, among many that exist, are shown in Figure 1.

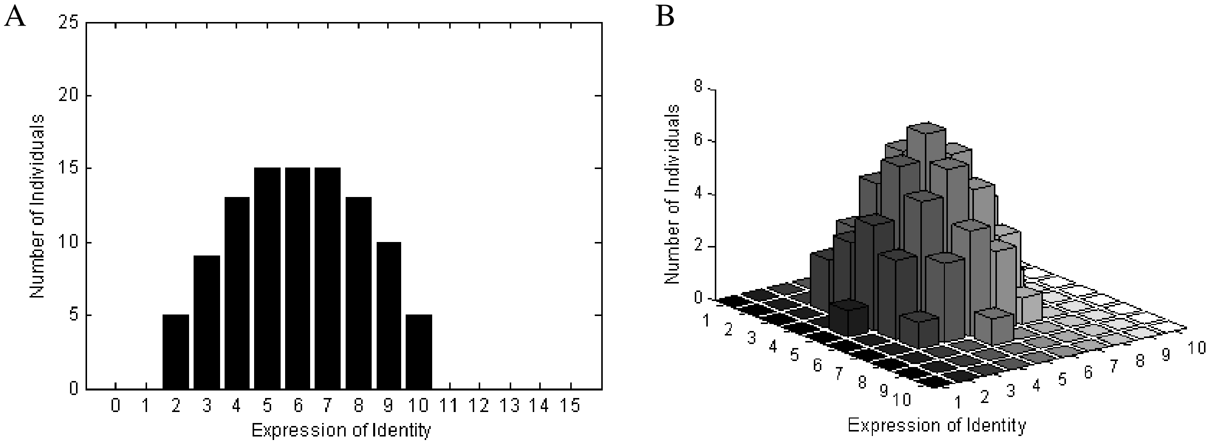

Figure 1: Two Nash equilibria distributions of identity expression for populations of $N = 100$ individuals. We set $\lambda = 1.5$ for this illustration. (**A**): Expression of a single one-dimensional trait over the domain $\{0..15\}$. (**B**): Expression of a single two-dimensional trait over the domain $\{1..10\}^2$. By symmetry, the distributions can be shifted anywhere within these (or wider) domains, and many strategy profiles give rise to the same population distributions. Even after accounting for these symmetries, these Nash equilibria are not unique.

Theorem 1 says that in a well-mixed population, in the long run we will not see popularity cycles, perpetual change, or novelty. The fact that we do, in reality, observe popularity cycles, perpetual change, and novelty suggests that we should consider a more realistic model. We now consider the social dynamics that result from assuming that people care only about the expressed identities of their immediate neighbors in their social network.

Model 2: Social Identity Expression in Social Networks

A social network is described by an adjacency matrix A where $a_{ij} = 1$ if person i observes, and thus cares about, person j's expressed identity (and equals 0 if not). Let $\eta(i) = \{j : a_{ij} = 1\}$ denote the set of people that person i observes, i.e., his neighbors.

Conformity among one's neighbors depends on distance from one's neighbors' average identity, $\bar{x}_{\eta(i)}$. Uniqueness among one's neighbors depends on the average amount of shared traits among one's neighbors (or, more precisely, the average across the different aspects of identity of the number of neighbors who express the same trait as oneself), denoted $\tilde{n}_i(X; \eta(i))$. Thus, we now model person i's utility given the profile of expressed identities X and his set of neighbors $\eta(i)$ as

$$u_i(X) = -\|x_i - \bar{x}_{\eta(i)}\|^2 - \lambda\, \tilde{n}_i(X; \eta(i)). \tag{2}$$

Results: Social Dynamics in Social Networks

Theorem 2. *Suppose people derive utility from both their conformity and their uniqueness among their neighbors in a social network, as in Equation (2) with $\lambda > 1$ and $m = 1$. Then there exists a social network adjacency matrix $\hat{A}$ such that no pure strategy Nash equilibrium exists and, thus, better-reply dynamics never converge to an absorbing state.*

Proof. By construction. We provide an example of a social network with $N = 3$ people that illustrates the result. (Any larger social network that contains this network as an out-component also suffices.) Let person 1 observe (only) person 2, person 2 observe (only) person 3, and person 3 observe (only) person 1.

Observe that the best response correspondence for each person is as follows:

$$x_1^* \in \{x : \|x - x_2\|^2 = 1\}$$
$$x_2^* \in \{x : \|x - x_3\|^2 = 1\}$$
$$x_3^* \in \{x : \|x - x_1\|^2 = 1\}.$$

Each person wants to be one unit of distance away from the person he is observing. If we associate the parity of an expressed identity x with two colors (i.e., distinguish only whether the sum of its integer coordinates is even or odd), then each person wants to have the color different from the person he is observing. However, it is impossible for all three people to simultaneously choose best responses because of the mathematical fact that odd-length cycle graphs are not 2-colorable. □

Theorem 2 says that with only local interactions in a social network, perpetually changing identity expression and popularity cycles become possible. Observe that the uniqueness motive is critical for obtaining this result. If we were to eliminate the uniqueness motive by setting $\lambda = 0$, then any homogeneous profile of expressed identities (with x_i identical for all i) would be a pure strategy Nash equilibrium, regardless of the social network structure. The uniqueness motive along with the local interactions together allow for more realistic, complex social dynamics.

Still, Theorem 2 only provides an existence result constructed with a highly stylized, simplistic social network. It does not tell us whether complex social dynamics typically emerge from our model when people are connected by realistic social networks. Real social networks have community structure with high levels of triadic closure (i.e., clustering or transitivity) – people associate mostly in small, tightly knit groups (Granovetter, 1973; Girvan and Newman, 2002; Newman and Park, 2003). This community structure does not typically include the kind of isolated cycle invoked in the proof of Theorem 2. We now use computational modeling to explore the dynamics of our model on realistic social networks.

Realistic Social Networks

We used a variant of the Jin-Girvan-Newman algorithm (Jin et al., 2001) to create a sample of 25 directed social networks with positive levels of clustering and community structure and limited out-degree. The networks have $N = 100$ people, each of whom can observe up to a maximum of $z_{\max}$ neighbors. Connections are formed and broken randomly, with a tendency to begin observing specific individuals who currently either observe or are observed by others who one is already observing. (Real social networks exhibit both patterns of directed closure (Brzozowski and Romero, 2011).) This tendency for clustering depends on a free parameter r. We varied r in $\{.01, .05, .1, .5, 1\}$ and $z_{\max}$ in $\{3..7\}$ to create the 25 networks. (See *Materials and Methods* for additional details.) Networks with higher $z_{\max}$ have more connections, and networks with higher r are more tightly clustered.

For each of these social networks, we repeatedly computed better-reply dynamics, specified with a simple random search for better replies based on the utility function in Equation (2) with λ in $\{0.5, 1.5, 5.0\}$, to see how often the dynamics converged to equilibrium within $1,000,000$ time steps. (Different specifications of better-reply dynamics could lead to different patterns of identity expression, but they all share the property that their rest points are the Nash equilibria of the game, so our results should be robust across this class of dynamics.) For robustness we considered three different specifications of the space of possible identities: first, $m = 1$, $d = 1$, and $\{a..b\} = \{0..99\}$; second, $m = 1$, $d = 2$, and $\{a..b\} = \{0..9\}$; and third, $m = 2$, $d = 1$, and $\{a..b\} = \{0..9\}$. (Higher dimensional spaces for identity expression would be more realistic, but are too computationally intensive to explore. We simply made the spaces large enough that everybody could express unique identities.) We repeated each computation 10 times, for a total of 2250 trials across the 9 different parameter specifications and 25 networks. (See *Materials and Methods* for additional details.) If the dynamics did not converge within $1,000,000$ time steps, we classified them as non-convergent (for that trial). (We believe the cutoff at $1,000,000$ time steps provides ample time for convergence, because we first computed the dynamics in the full, well-mixed population, for which Theorem 1 tells us that they must converge, and found that across 90

trials, the dynamics always converged within 2000 time steps. We discuss additional checks on the sufficiency of $1,000,000$ time steps below.)

Computational Results: Frequency of Non-Convergence

The frequency of non-convergent trials varied with the parameters specifying the game and the network formation process, with the value of λ in particular playing a critical role. When $\lambda = 0.5$, the dynamics usually converged to equilibrium (68.53% of these 750 trials). Figure 2A shows the frequency of convergent trials for each of the 25 networks, for each of the three specifications of the space of identities, with $\lambda = 0.5$. Darker shading indicates higher frequencies of convergence. The frequency of convergence varies non-monotonically with the maximum out-degree of the network $z_{\max}$. For $z_{\max} = 3$ or 4, the dynamics almost always converge, whereas for $z_{\max} = 7$, the dynamics usually do not converge. Yet there is more convergence with $z_{\max} = 6$ than with $z_{\max} = 5$.

When $\lambda = 1.5$, the dynamics usually did not converge (only in 18% of these 750 trials). Figure 2B shows the frequency of convergent trials for each of the 25 networks, for each of the three specifications of the space of identities, with $\lambda = 1.5$. Four of the networks with $z_{\max} = 4$ usually converged (specifically, those with $r > .01$). A few of the other networks occasionally converged. Many never converged at all.

When $\lambda = 5$, the dynamics almost never converged. The only exception was the network with $z_{\max} = 4$ and $r = 1$, which converged in all 10 trials with $m = 1$ and $d = 1$. However, none of the other 740 trials with $\lambda = 5$ converged.

The results presented here leave room for two arguments raising concern that perhaps the dynamics would always eventually converge if they just had more time to continue running. First, it is surprising to see so many parameter specifications for which the dynamics sometimes converge and other times do not. We might have expected non-convergent trials whenever there is no pure Nash equilibrium, but that whenever such an equilibrium exists and convergence is possible, it would eventually occur. Perhaps it just needs more time. However, even when a pure Nash equi-

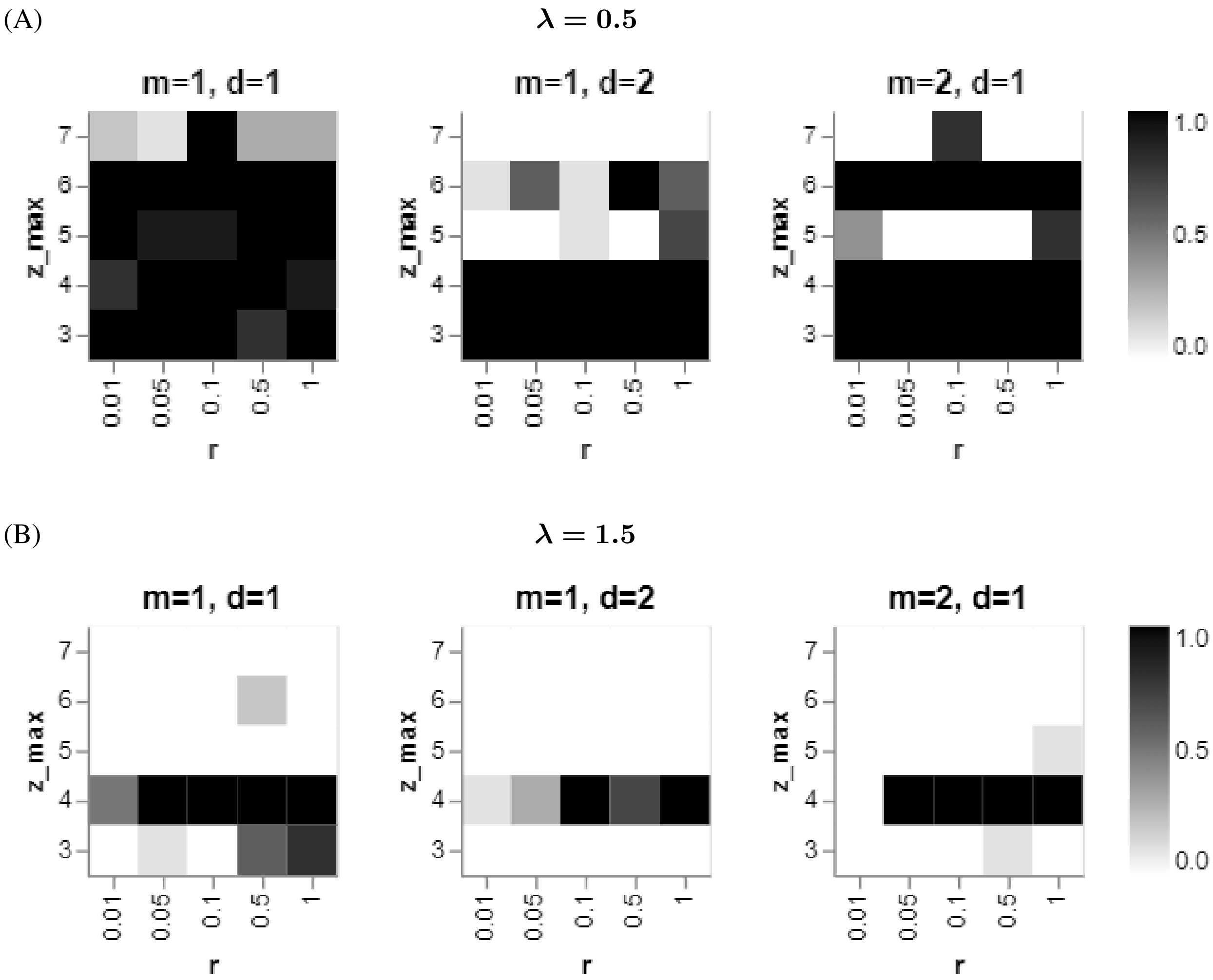

Figure 2: Frequency of convergent trials for each network. Darker shading indicates higher frequencies of convergence. The trials with $\lambda = 5$ are omitted because they almost never converged.

librium does exist, allowing the dynamics to converge in some trials, it is possible for the dynamics to enter a random walk on an absorbing subspace, from which it is no longer possible to reach the equilibrium. This could explain the observed frequencies of convergence that are positive but still less than 100%. Still, a second cause for concern is that larger values of λ give the better-reply dynamics more possible states to explore when a neighbor adopts one's own identity. Thus, we should expect it to take longer to reach an equilibrium with larger values of λ. If the dynamics usually converge with $\lambda = 0.5$, might they be on their way, but not quite there yet, with larger values of λ?

A few additional pieces of data reassure us that most of the trials we have classified as non-

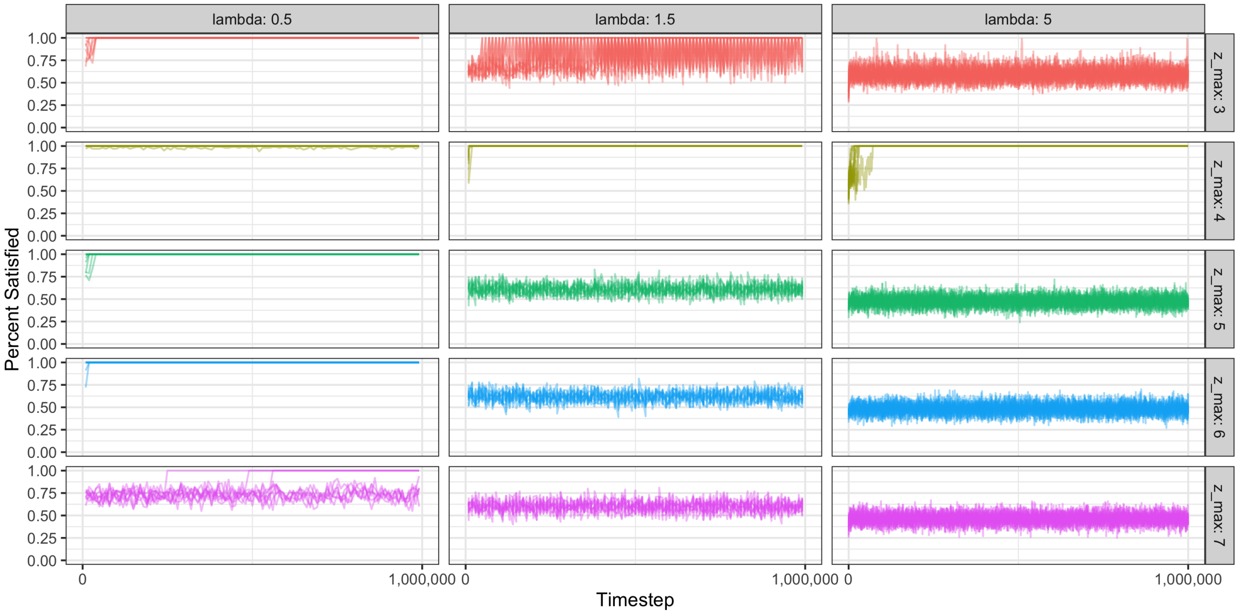

Figure 3: Percentage of individuals satisfied over 1,000,000 time steps for each trial with $m = 1, d = 1$, and varying λ, for networks with $r = 1$ and varying $z_{\max}$.

convergent are not artifacts of terminating the computation too quickly. First, for each trial we examine the fraction of individuals that are satisfied with their current identities every 1000 time steps during the trial. Convergence to equilibrium occurs if and when everybody is satisfied. So, the trajectories of the percentage of satisfied individuals also reveal the times to reach equilibrium, when convergence occurs. Figure 3 shows the percentage of satisfied individuals over time for each trial with $m = 1$, $d = 1$, and varying λ, for networks with $r = 1$. Figures SM1 and SM2 in the Supplemental Materials show the corresponding results with $m = 1, d = 2$ and with $m = 2, d = 1$ respectively. The results for networks with $r < 1$ look similar and are omitted. Across the board, when the dynamics do converge to equilibrium, they tend to do so quickly. Although the distribution of convergence times does have a fat tail, it certainly appears that convergence becomes less and less likely over time. Additionally, while the percentage of satisfied individuals appears to bounce around randomly, for many of the parameter values it appears to be bounded well below 100%.

The trajectories of the percentage of satisfied individuals suggest that the trials we have deemed non-convergent really would never converge, but of course there can be no guarantee. With

$N = 100$ individuals choosing among 100 possible identities, it is simply not computationally feasible to check every possible scenario. However, with $N = 8$ individuals choosing among 8 possible identities, it is feasible to exhaustively search for equilibria. We created an additional social network using the same algorithm with $z_{\max} = 3$ and $r = 1$, but with $N = 8$. Once again, the better-reply dynamics with $\lambda = 5$ and $m = 1$, $d = 1$, and $\{a..b\} = \{0..7\}$ did not converge. We then exhaustively searched every profile of identities on this space and verified that no pure Nash equilibrium exists. This guarantees that the dynamics would never converge. This network does not contain an isolated odd-cycle, which our proof of Theorem 2 relied on, but it provides another example that shows that non-convergence is possible, and moreover can occur with realistic network structure.

We interpret these results to mean that when the uniqueness motive is sufficiently strong, the dynamics on realistic social networks usually will not converge. However, if the uniqueness motive is too weak, individuals feel little pressure to differentiate themselves, and they may settle into an equilibrium with overlapping identities.

Computational Results: Conformity

We further explore the dynamics by observing the trajectories of identity expression over the initial $10,000$ time steps. Clearly, because of the uniqueness motive, there will always be some diversity of identity expression. As the uniqueness motive gets stronger, i.e., as λ increases, we expect to observe less conformity. Sure enough, this is the case. Figure 4 displays the distributions of the distances from individuals' identities to the average identity in the population and to the average identity of their neighbors in the network, $\|x_i - \bar{x}\|$ and $\|x_i - \bar{x}_{\eta(i)}\|$ respectively, measured at the $10,000$th time step, for $m = 1$, $d = 1$, and varying λ, aggregating trials across the different networks. Figures SM3 and SM4 in the Supplemental Materials show the corresponding results for $m = 1, d = 2$ and for $m = 2, d = 1$ respectively.

We first compare the average distance to the population mean expressed identity across different values of λ. The average distance to the population mean increased from 0.59 (SD $= 0.57$)

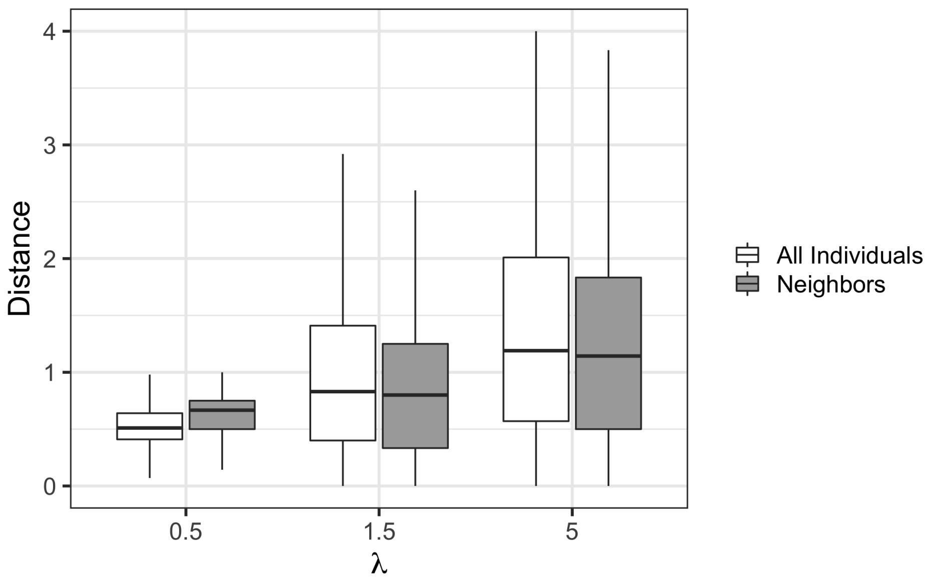

Figure 4: Box plots showing distances from individuals' identities to the average identity of all individuals in the population and to the average identity of their neighbors in the network, measured at the $10,000$th time step, for $m = 1$, $d = 1$, and varying λ, aggregating trials across the different networks.

when $\lambda = 0.5$ to 0.99 (SD $= 0.83$) when $\lambda = 1.5$ to 1.45 (SD $= 1.13$) when $\lambda = 5$. Both of these increases were statistically significant with $p < .001$ in t-tests ($t(440040) = -195.69$ for the comparison between distances when $\lambda = 0.5$ and $\lambda = 1.5$, and $t(460375) = -162.82$ for the comparison between distances when $\lambda = 1.5$ and $\lambda = 5$). We then compare the average distance to one's neighbors across different values of λ. The average distance to one's neighbors increased from 0.61 (SD $= 0.26$) when $\lambda = 0.5$ to 0.85 (SD $= 0.58$) when $\lambda = 1.5$ to 1.25 (SD $= 0.86$) when $\lambda = 5$. Again, both of these increases were statistically significant with $p < .001$ in t-tests ($t(349573) = -191.64$ for the first, and $t(436957) = -191.08$ for the second).

We also check whether the expressed identities display the signature empirical pattern associated with social influence: do individuals express identities that are more similar to their network neighbors' identities than to the average member of the population as a whole? The differences between the distances to the population mean identity and to the mean of one's neighbors' identities appear to be small in Figure 4, but they are all statistically significant with $p < .001$ in paired t-tests ($t(249799) = -22.28$ for the comparison when $\lambda = 0.5$, $t(249799) = 91.14$ for the comparison when $\lambda = 1.5$, and $t(249799) = 93.44$ for the comparison when $\lambda = 5$). For $\lambda = 1.5$ and $\lambda = 5$, individuals do indeed express identities that more closely resemble the people they observe than others in the population. Yet for $\lambda = 0.5$, individuals are actually more similar to unobserved others than to their network neighbors. There is little diversity across the entire population in these trials.

Computational Results: Momentum and Contagion

Next we look for momentum in the dynamics. For simplicity, we restrict this analysis to trials with $m = 1$ and $d = 1$. As a measure of momentum over time, we compute $\sigma_{100}(t) = \frac{1}{100}\sum_{t'=1}^{100} \Delta x(t) * \Delta x(t + t')$, where $\Delta x(t)$ is the change in identity expression of the individual who searched for a better reply at time step t. We take the average momentum for a trial to be the average value of $\sigma_{100}(t)$ for $1000 \leq t < 9900$. (We exclude the first 1000 time steps because they tend to be noisy.) Figure 5 shows the average momentum on each network for varying λ, aggregated over 10 trials. We observe that average momentum is always positive, and a t-test shows it to be significantly different from zero (M $=$ .0096, SD $=$ .31), $t(6674999) = 80.87$, $p < .001$, indicating that changes in identity expression tend to persist in the same direction over time.

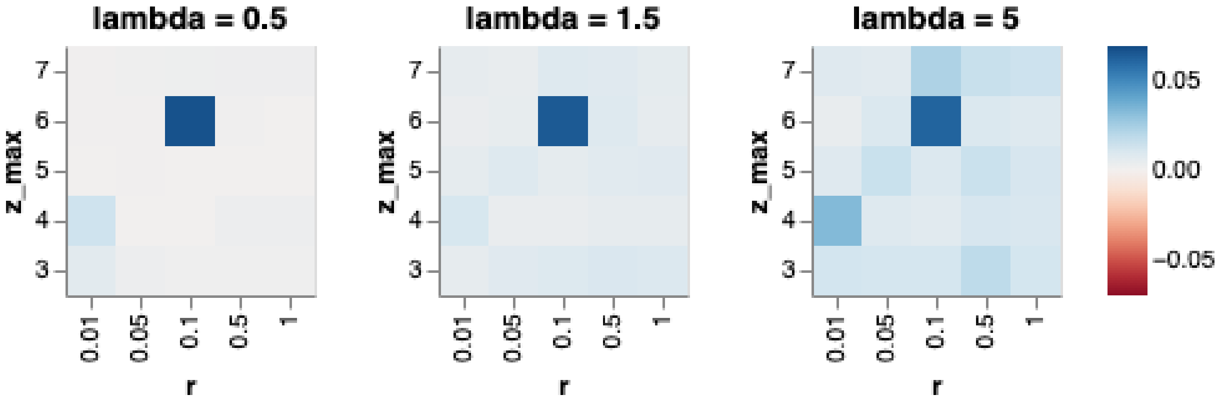

Figure 5: Average momentum on each network for varying λ, with $m = 1$ and $d = 1$, aggregated over 10 trials. In all cases, the average momentum is positive. Darker shading indicates greater momentum.

Figure 5 also shows clear differences in the average momentum across the different networks.

Most prominently, we observe particularly strong momentum on the network with $r = 0.1$ and $z_{\max} = 6$. This finding is robust across multiple trials, not the result of a single outlying trial, but appears to be specific to this particular network. (We created another network with the same parameters, $r = 0.1$ and $z_{\max} = 6$, to see if this result would replicate. It did not. In the attempted replication, the average momentum aggregated over 30 trials across the same λ values was $.006$.) We examined the network's properties (available in the SM) hoping to explain why strong momentum develops on this network, but the network does not appear to have unusual characteristics or structure.

We use multiple linear regression to assess how momentum depends on our parameters r, $z_{\max}$ and λ. Table 1 reports the results. We find that average momentum is increasing in λ and $z_{\max}$. Intuitively, higher values of λ make individuals willing to make larger shifts in their identity to remain unique, which generates stronger momentum, and higher values of $z_{\max}$ mean that a single person's change in identity affects more of the other people in the network who observe that change, which also generates stronger momentum.

Table 1: Linear regression of average momentum.

Effect	Estimate	SE	p
λ	.0051	.0001	$< .001$
$z_{\max}$	.0007	.0001	$< .001$
r	$-.0001$	.0001	.291
Constant	$-.002$	.0005	$< .001$
Observations	$6,675,000$		
R^2	.0002		
Adjusted R^2	.0002		
Residual Std. Error	.3062	(df = $6,674,996$)	
F Statistic	430.8	(df = $3;\ 6,674,996$)	$p < .001$

We were particularly interested in how average momentum depends on r, because this distinguishes a complex contagion from a simple contagion. In a simple contagion, there would be greater momentum when there is less clustering (smaller r), whereas in a complex contagion, there would be greater momentum when there is more clustering (larger r). However, we find no significant linear trend here. Qualitatively, it appears that momentum is strongest for an intermediate level of clustering (perhaps generating bridges that are both long and wide), but this speculative finding might just reflect the observation of particularly strong momentum on the single network with $r = 0.1$ and $z_{\max} = 6$.

Model 3: Co-Evolution of Social Identity Expression and Social Networks

Up to this point, we have considered social identity expression on fixed social networks, but social networks themselves evolve over time. There is ample empirical evidence that people are more likely to form (and less likely to dissolve) all kinds of relationships with people who are more similar to them – a pattern of social network dynamics known as homophily (McPherson et al., 2001). By first forming the social networks and then considering the dynamics of social identity expression on these fixed networks, we could capture a form of social influence, but we could not capture homophily. We now consider integrating the dynamics of social identity expression with the dynamics of social network formation, to incorporate homophily. We investigate whether our earlier results are robust in this model of co-evolving identities and social network ties.

The model relies on the same utility function, given in Equation 2. Now, at each time step an individual can either consider a change in his own identity or a change in the network neighbors he observes. (We assume each consideration is equally likely.) In the former case, the individual randomly considers a new expression of identity. In the latter case, the individual considers forming a new connection either to a randomly selected other person or specifically to another person who already has a link (in either direction) with someone he already has a connection to (i.e., with a tendency toward triadic closure), and if the focal individual already had as many relationships as he could handle, he simultaneously considers breaking an existing connection. (In reality, limits on the number of relationships an individual can handle are likely to be somewhat more flexible, but this stylized model parsimoniously captures the clustering and bounded out-degree that characterize social networks.) Critically, the individual only accepts changes to his identity or to his social network if they increase his utility. (An exception is made for the first network connection that each individual considers forming, which is always accepted, because the utility function is not well defined if the individual has no connections at all.) See *Materials and Methods* for additional details about the process. The model effectively brings together Jin et al.'s (2001) social network formation process with the preferences about social identity expression that we have proposed here, and induces homophily by only allowing changes to one's social network that increase utility.

We ran 30 trials each with $\lambda = 0.5$, $\lambda = 1.5$, and $\lambda = 5$. When $\lambda = 0.5$, 23% $(7/30)$ of the trials converged to equilibrium. When $\lambda = 1.5$ or $\lambda = 5$, none of the trials converged to equilibrium. These results are consistent with our earlier results for the fixed social networks.

We again compare the conformity among network neighbors to the conformity in the population as a whole. Figure 6 shows the average distances to the population mean identity and to one's neighbors' mean identity, measured at the end of the trial, for each λ, aggregating across the 30 trials. We find that expressed identities are significantly more similar to one's neighbors' identities than to the population mean identity in all three cases. The differences here are much starker than than they were in the comparisons on the fixed social networks because the social networks that endogenously form here are not necessarily fully connected. When people sort themselves into non-overlapping social groups, the distance between the groups' mean identities tends to be larger than the variance of identities within a group.

We again look for momentum in the dynamics. This time we simply compute the percentage of successive changes to identities that are in the same direction over the duration of each trial. Averaging across the trials, well above half (59%, 95% CI $[58.2\%, 59.9\%]$) of shifts in identity are in the same direction as the previous one. When we restrict to changes in identity within the largest connected component of the network after the first $10,000$ time steps, it jumps to almost always (99.4% of successive shifts, 95% CI $[99.37\%, 99.45\%]$) going in the same direction. Thus, the finding of significant momentum carries through from our earlier results for the fixed social networks.

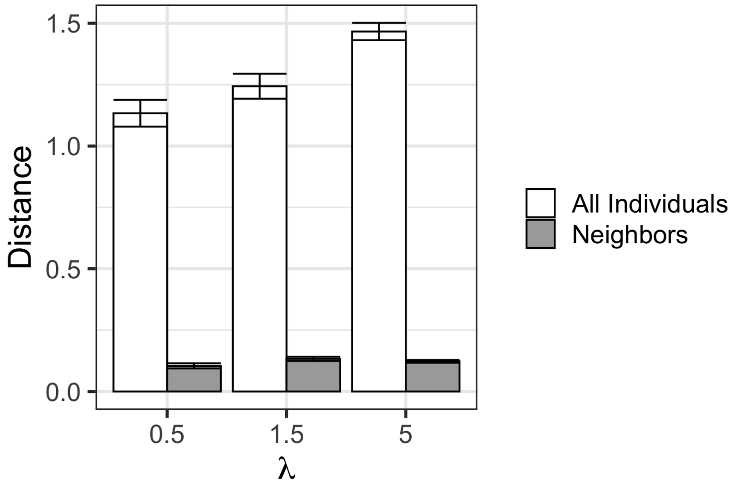

Figure 6: Average distance to the population mean identity and to one's neighbors' mean identity, measured at the end of the trial, for each λ, aggregating across the 30 trials.

Discussion

These results tell us that with local interactions on realistic social networks, the interplay of conformity and sufficiently strong uniqueness motives produces social dynamics for identity expression that are indeed typically non-convergent. People continually change their expressed identities, and certain forms of expression come into and out of fashion in unpredictable cycles. Popularity cycles are inherently unpredictable in the model because people typically have multiple better replies (and even multiple best responses) to choose from in the face of most profiles of their neighbors' identity expression. The multiplicity of paths the dynamics could take leaves room for idiosyncrasy.

Our findings help us understand the role of social networks and local interaction in the dynamics of cultural trends. Popularity cycles, perpetual change, and novel expressions of social identity should be expected when people observe their neighbors in realistic, directed social networks and care about being unique as well as fitting in. While popularity cycles are often attributed to chase-and-flight dynamics arising from asymmetric imitation and differentiation, complex social dynamics of identity expression may also arise from our alternative specification of conformity and uniqueness preferences and social network structure.

Recognition of conformity and uniqueness as opposing, but not mutually exclusive, motives is also part of optimal distinctiveness theory (Brewer, 1991; Leonardelli et al., 2010). However, optimal distinctiveness theory posits that people form collective identities by choosing to associate themselves with social groups, whereas our concept of social identity operates at the level of the individual. In our view, collective identities emerge at the level of the group based on their members' individual identities. From the alternative, similarly valid perspective, we could propose that individual identities emerge from a psychological process of finding consonance between the collective identities of the many groups that an individual affiliates with at any point in time. Connecting these perspectives requires deeper understanding of how people choose to associate with or withdraw from social groups, and how this relates to social network structure. While this integration remains beyond our present grasp, we find it useful to have complementary theories aimed at different levels of social identity.

We use game theory and computational modeling here to describe social dynamics with mathematical precision. Social phenomena do not always reflect individual preferences (Schelling, 1969, 1971). Mathematical modeling helps us understand the relationship between individual motives and aggregate social dynamics when interactions generate nontrivial feedbacks. Our work here is part of a tradition of formal modeling of social identity and fashion (Miller et al., 1993; Strang and Macy, 2001; Tassier, 2004; Smaldino et al., 2012; Smaldino and Epstein, 2015; Smaldino et al., 2015; Brown et al., 2019). This approach yields us deep theoretical insight, and we hope it inspires more research leading to further insights into social dynamics and identity expression.

Materials and Methods

The Social Networks

We borrow Jin, Girvan, and Newman's Model II algorithm for growing undirected social networks (Jin et al., 2001) and modify it to generate directed social networks with $N = 100$ people, each of whom can observe up to a maximum of $z_{\max}$ neighbors. The network is initialized with all 100 people and no connections. The following three steps are then repeated 100 times:

1. Choose 3 pairs of individuals uniformly at random. For each pair i and j, if i observes less than $z_{\max}$ people and does not already observe j, then i begins to observe j; else, if j observes less than $z_{\max}$ people and does not already observe i, then j begins to observe i.

2. Randomly select a fraction r of the triads i, j, and k such that i observes k and k observes j or that i and j both observe k. If i observes less than $z_{\max}$ people and does not already observe j, then i begins to observe j.

3. Randomly select and break 0.5% of connections (rounded up).

All 25 social networks, measures of their structural properties, and the Python source code used to create them are made available in the SM Appendix.

The Better-Reply Dynamics

Our computational model adopts a specification of the better-reply dynamics in which at each time step, one randomly selected individual searches for (and upon discovery, adopts) a better reply to the current population profile. Initial strategies are randomly (uniformly) distributed. We check for convergence every 1000 time steps by checking whether any individual can find a better reply. The Python source code and complete output data are available in the SM Appendix.

Co-evolving Social Networks and Identities

We again assume there are $N = 100$ people. We consider the space of identities with $m = 1$, $d = 1$, and $\{a..b\} = \{0..9\}$. We set the maximum number of neighbors that an individual can handle (i.e., maximum out-degree) to be $z_{\max} = 5$. In this model, in contrast to the earlier model, each time step corresponds to a single individual considering a single change (either to his identity or his network), rather than searching for (i.e., repeatedly considering) such a change. We allow the dynamics to run for up to $2,000,000$ time steps before cutting them off and classifying them as non-convergent, and we check for convergence every 1000 time steps.

Initially people have no network connections and strategies are randomly distributed. At each time step, there is an equal 50% chance of considering a change in identity or a change in the network. In the former case, a randomly selected individual considers switching to a randomly selected new identity and does so only if the switch increases his utility. In the latter case, the probability of considering a new connection from person i to person j is proportional to $1 + 2000(\tau_{\text{in}} + \tau_{\text{out}})$, where τ_{in} is the number of triads in which i and j both observe some other individual k, and τ_{out} is the number of triads in which i observes some other individual k, who then observes j. If person i already has $z_{\max}$ connections to other people, then the potential connection to j is considered jointly with breaking one of i's existing connections. Person i goes through with the change only if it would increase his utility or if he previously had no connections (in which case his utility was not yet well defined). The Python source code and output data are made available in the SM appendix.

Disclosures and Acknowledgments

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests: The authors declare that they have no competing financial interests.

Acknowledgements: Thanks to David Zixi Chen, Andrew Gu, Cici Yidan Hu, Ashwin Konkimalla, Tzen-Hao Ng, and Andrew Wu for research assistance.

Prior Dissemination: An earlier version of this paper was presented at the Economic Research on Identity, Norms, and Narratives Conference in 2018 and was posted to arXiv.

References

Abrahamson, Eric. 1991. "Managerial Fads and Fashions: the Diffusion and Refection of Innovations." *Academy of Management Review* 16:586–612.

Acerbi, Alberto and R Alexander Bentley. 2014. "Biases in cultural transmission shape the turnover of popular traits." *Evolution and Human Behavior* 35:228–236.

Acerbi, Alberto, Stefano Ghirlanda, and Magnus Enquist. 2012. "The Logic of Fashion Cycles." *PLoS One* 7:e32541.

Asch, Solomon E. 1955. "Opinions and Social Pressure." *Scientific American* 193:31–35.

Asch, Solomon E. 1956. "Studies of Independence and Conformity: I . A Minority of One Against a Unanimous Majority." *Psychological Monographs: General and Applied* 70:1–70.

Bakshi, Nitin, Kartik Hosanagar, and Christophe Van den Bulte. 2013. "Chase and flight: New product diffusion with social attraction and repulsion." *History* pp. 1–31.

Barucca, Paolo, Jacopo Rocchi, Enzo Marinari, Giorgio Parisi, and Federico Ricci-Tersenghi. 2015. "Cross-correlations of American baby names." *Proceedings of the National Academy of Sciences* 112:7943–7947.

Bentley, R Alexander, Matthew W Hahn, and Stephen J Shennan. 2004. "Random drift and culture change." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 271:1443–1450.

Bentley, R Alexander, Carl P Lipo, Harold A Herzog, and Matthew W Hahn. 2007. "Regular rates of popular culture change reflect random copying." *Evolution and Human Behavior* 28:151–158.

Berger, Jonah. 2008. "Identity signaling, social influence, and social contagion." In *Understanding peer influence in children and adolescents.*, edited by Mitchell J. Prinstein and Kenneth A. Dodge, pp. 181–199. New York: The Guilford Press.

Berger, Jonah, Eric T Bradlow, Alex Braunstein, and Yao Zhang. 2012. "From Karen to Katie: Using baby names to understand cultural evolution." *Psychological Science* 23:1067–1073.

Berger, Jonah and Chip Heath. 2007. "Where Consumers Diverge from Others: Identity Signaling and Product Domains." *Journal of Consumer Research* 34:121–134.

Berger, Jonah and Chip Heath. 2008. "Who drives divergence? Identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes." *Journal of personality and social psychology* 95:593.

Berger, Jonah and Gaël Le Mens. 2009. "How adoption speed affects the abandonment of cultural tastes." *Proceedings of the National Academy of Sciences* 106:8146–8150.

Brewer, Marilynn B. 1991. "The Social Self: On Being the Same and Different at the Same Time." *Personality and Social Psychology Bulletin* 17:475–482.

Brown, Gordon D.A., Stephan Lewandowsky, and Zhihong Huang. 2019. "Social Sampling Theory: Authenticity Preference and Social Extremeness Aversion Lead to Social Norm Effects and Polarization." .

Brzozowski, Michael J and Daniel M Romero. 2011. "Who Should I Follow? Recommending People in Directed Social Networks." In *ICWSM*, pp. 458–461.

Centola, Damon. 2010. "The spread of behavior in an online social network experiment." *Science* 329:1194–1197.

Centola, Damon and Michael Macy. 2007. "Complex contagions and the weakness of long ties." *American Journal of Sociology* 113:702–734.

Chan, Cindy, Jonah Berger, and Leaf Van Boven. 2012. "Identifiable but Not Identical: Combining Social Identity and Uniqueness Motives in Choice." *Journal of Consumer Research* 39:561–573.

Christakis, Nicholas A and James H Fowler. 2013. "Social contagion theory: examining dynamic social networks and human behavior." *Statistics in Medicine* 32:556–577.

Cialdini, Robert B. and Melanie R. Trost. 1998. "Social influence: Social norms, conformity and compliance." In *The handbook of social psychology, Vols. 1 and 2*, edited by D.T. Gilbert, S.T. Fiske, and G. Lindzey, pp. 151–192. Boston: McGraw-Hill.

Fiske, Susan T and Shelley E Taylor. 2013. *Social cognition: From brains to culture*. Sage.

Friedman, James W. and Claudio Mezzetti. 2001. "Learning in Games by Random Sampling." *Journal of Economic Theory* 98:55–84.

Gigerenzer, Gerd. 2000. *Adaptive thinking: Rationality in the real world*. Oxford University Press, USA.

Girvan, Michelle and Mark EJ Newman. 2002. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99:7821–7826.

Golman, Russell, George Loewenstein, Karl Ove Moene, and Luca Zarri. 2016. "The Preference for Belief Consonance." *Journal of Economic Perspectives* 30:165–188.

Granovetter, Mark S. 1973. "The strength of weak ties." *American journal of sociology* 78:1360–1380.

Gureckis, Todd M and Robert L Goldstone. 2009. "How you named your child: Understanding the relationship between individual decision making and collective outcomes." *Topics in Cognitive Science* 1:651–674.

Hetherington, Kevin. 1998. *Expressions of Identity: Space, Performance, Politics*. London: Sage Publications.

Hofbauer, Josef and Karl Sigmund. 2003. "Evolutionary game dynamics." *Bulletin of the American mathematical society* 40:479–519.

Imhoff, Roland and Hans-Peter Erb. 2009. "What Motivates Nonconformity? Uniqueness Seeking Blocks Majority Influence." *Personality and Social Psychology Bulletin* 35:309–320.

Jin, Emily M, Michelle Girvan, and Mark E. J. Newman. 2001. "Structure of growing social networks." *Physical Review E* 64:046132.

Karni, Edi and David Schmeidler. 1990. “Fixed Preferences and Changing Tastes.” *The American Economic Review, Papers and Proceedings* 80:262–267.

Keynes, John Maynard. 1936. *General theory of employment, interest and money*. New York: Harcourt Brace.

Leibenstein, Harvey. 1950. “Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand.” *The Quarterly Journal of Economics* 64:183–207.

Leonardelli, Geoffrey J., Cynthia L. Pickett, and Marilynn B. Brewer. 2010. “Optimal Distinctiveness Theory: A Framework for Social Identity, Social Cognition, and Intergroup Relations.” *Advances in Experimental Social Psychology* 43:63–113.

Lieberson, Stanley. 2000. *A Matter of Taste: How Names, Fashions, and Culture Change*. New Haven: Yale University Press.

Lieberson, Stanley and Freda B. Lynn. 2003. “Popularity as taste: an application to the naming process.” *Onoma* 38:235–276.

Lynn, Michael and Charles R. Snyder. 2002. “Uniqueness.” In *Handbook of positive psychology*, edited by C. R. Snyder and Shane J. Lopez, pp. 395–410. New York: Oxford University Press.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. “Birds of a feather: Homophily in social networks.” *Annual review of sociology* 27:415–444.

Miller, Christopher M, Shelby H McIntyre, and Murali K Mantrala. 1993. “Toward Formalizing Fashion Theory.” *Journal of Marketing Research* 30:142–157.

Monderer, Dov and Lloyd S. Shapley. 1996a. “Fictitious Play Property for Games with Identical Interests.” *Journal of Economic Theory* 68:258–265.

Monderer, Dov and Lloyd S. Shapley. 1996b. “Potential Games.” *Games and Economic Behavior* 14:124–143.

Newman, Mark EJ and Juyong Park. 2003. “Why social networks are different from other types of networks.” *Physical Review E* 68:036122.

Pesendorfer, Wolfgang. 1995. “Design Innovation and Fashion Cycles.” *American Economic Review* 85:771–792.

Rentfrow, Peter J. and Samuel D. Gosling. 2006. “Message in a ballad: the roles of music preferences in interperosnal perception.” *Psychological Science* 17:236–242.

Reynolds, William H. 1968. “Cars and Clothing: Understanding Fashion Trends.” *Journal of Marketing* 32:44–49.

Richardson, Jane and Alfred Louis Kroeber. 1940. “Three centuries of women’s dress fashions, a quantitative analysis.” *Anthropological Records* 5:111–153.

Rosenthal, Robert W. 1973. “A class of games possessing pure-strategy Nash equilibria.” *International Journal of Game Theory* 2:65–67.

Salganik, Matthew J, Peter Sheridan Dodds, and Duncan J Watts. 2006. “Experimental study of inequality and unpredictability in an artificial cultural market.” *Science* 311:854–856.

Salganik, Matthew J and Duncan J Watts. 2008. “Leading the Herd Astray: An Experimental Study of Self-fulfilling Prophecies in an Artificial Cultural Market.” *Social Psychology Quarterly* 71:338–355.

Schelling, Thomas C. 1969. “Models of Segregation.” *American Economic Review* 59:488–493.

Schelling, Thomas C. 1971. “Dynamic Models of Segregation.” *Journal of Mathematical Sociology* 1:143–186.

Shuker, Roy. 2016. *Understanding popular music culture*. New York: Routledge.

Simmel, Georg. 1957. “Fashion.” *American Journal of Sociology* 62:541–558.

Smaldino, Paul E, Jimmy Calanchini, and Cynthia L Pickett. 2015. "Theory development with agent-based models." *Organizational Psychology Review* 5:300–317.

Smaldino, Paul E and Joshua M Epstein. 2015. "Social conformity despite individual preferences for distinctiveness." *Royal Society Open Science* 2:140437–140437.

Smaldino, Paul E, Cynthia L Pickett, Jeffrey Sherman, and Jeffrey Schank. 2012. "An Agent-Based Model of Social Identity Dynamics." *Journal of Artificial Societies and Social Simulation* 15:1–17.

Snyder, Charles R. and Howard L. Fromkin. 1980. *Uniqueness: The Human Pursuit of Difference*. New York: Plenum Press.

Sproles, George B. 1981. "Analyzing Fashion Life Cycles: Principles and Perspectives." *Journal Of Marketing* 45:116–124.

Strang, David and Michael W Macy. 2001. "In Search of Excellence: Fads, Success Stories, and Adaptive Emulation." *American Journal of Sociology* 107:147–182.

Tassier, Troy. 2004. "A model of fads, fashions, and group formation." *Complexity* 9:51–61.

Turner, John C., Michael A. Hogg, Penelope J. Oakes, Stephen D. Reicher, and Margaret S. Wetherell. 1987. *Rediscovering the social group: A self-categorization theory.* Oxford: Basil Blackwell.

Zhang, Boyu, Zhigang Cao, Cheng-Zhong Qin, and Xiaoguang Yang. 2018. "Fashion and homophily." *Operations Research* 66:1486–1497.

Zuckerman, Ezra W. 2012. "Construction, Concentration, and (Dis)Continuities in Social Valuations." *Annual Review of Sociology* 38:223–245.

Supplemental Materials

Formal Definitions

We can express person i's neighbors' average identity as

$$\bar{x}_{\eta(i)} = \frac{1}{|\eta(i)|} \sum_{j \in \eta(i)} x_j.$$

We can express the number of i's neighbors who adopt the same expression of identity trait μ as person i as

$$\tilde{n}_{i,\mu}(X; \eta(i)) = \sum_{j \in \eta(i)} \delta(x_{i,\mu}, x_{j,\mu}),$$

where δ is the Kronecker delta function. Then

$$\tilde{n}_i(X; \eta(i)) = \frac{1}{m} \sum_{\mu} \tilde{n}_{i,\mu}(X; \eta(i))$$

is the average number of neighbors sharing one's traits (across all the aspects of identity). In a well-mixed population, we set $\eta(i) = \{j : j \neq i\}$ to recover $n_{i,\mu}(X)$ and $n_i(X)$ for all i.

Supplementary Results and Proofs

Lemma 1. *In a well-mixed population with utility functions given in Equation (1), the game has an exact potential function:*

$$\Phi(X) = -\sum_{i=1}^{N} \frac{N-1}{N} \|x_i - \bar{x}\|^2 + \frac{1}{2} \lambda \, n_i(X).$$

Proof. Consider a change in the profile of identities $X \to X'$ resulting from person i alone changing his identity $x_i \to x_i'$, i.e., such that $x_j' = x_j$ for all $j \neq i$. We need only show that the change in the potential function equals the change in i's utility: $\Phi(X') - \Phi(X) = u_i(X') - u_i(X)$.

We express the change in the potential function as a sum of the changes in each term:

$$\Phi(X') - \Phi(X) = \sum_{j=1}^{N} \frac{N-1}{N} \left(\|x_j - \bar{x}\|^2 - \|x'_j - \bar{x}'\|^2 \right) + \sum_{j=1}^{N} \frac{1}{2} \lambda \left(n_j(X) - n_j(X') \right).$$

We consider each of the two summations separately.

We expand the first sum:

$$\sum_{j=1}^{N} \frac{N-1}{N} \left(\|x_j - \bar{x}\|^2 - \|x'_j - \bar{x}'\|^2 \right) = \frac{N-1}{N} \left(\|x_i - \bar{x}\|^2 - \|x'_i - \bar{x}'\|^2 \right) + \sum_{j \neq i} \frac{N-1}{N} \left(\|x_j - \bar{x}\|^2 - \|x'_j - \bar{x}'\|^2 \right). \quad (3)$$

We find it useful to express the average identity as $\bar{x} = \frac{N-1}{N} \bar{x}_{-i} + \frac{1}{N} x_i$. Plugging in to the first term in Equation (3), we have:

$$\|x_i - \bar{x}\|^2 - \|x'_i - \bar{x}'\|^2 = \left(\frac{N-1}{N} \right)^2 \left(\|x_i - \bar{x}_{-i}\|^2 - \|x'_i - \bar{x}_{-i}\|^2 \right).$$

Plugging in to the second term in Equation (3), expanding and canceling off common terms, we have for any $j \neq i$:

$$\|x_j - \bar{x}\|^2 - \|x'_j - \bar{x}'\|^2 = \frac{1}{N^2} \left(\|x_i - \bar{x}_{-i}\|^2 - \|x'_i - \bar{x}_{-i}\|^2 \right) + \frac{2}{N} (x_j - \bar{x}_{-i}) \cdot (x_i - x'_i).$$

Observe that the last term here drops out when we sum over all $j \neq i$ because $\sum_{j \neq i} (x_j - \bar{x}_{-i}) = 0$. The first term does not depend on j, so summing over all $j \neq i$ just multiplies this term by a factor of $(N-1)$. Putting it all together, we find that Equation (3) simplifies to:

$$\begin{aligned}\sum_{j=1}^{N} \frac{N-1}{N} \left(\|x_j - \bar{x}\|^2 - \|x_j' - \bar{x}'\|^2\right) \\ = \left(\frac{(N-1)^3}{N^3} + \frac{(N-1)^2}{N^3}\right) \left(\|x_i - \bar{x}_{-i}\|^2 - \|x_i' - \bar{x}_{-i}\|^2\right) \\ = \left(\frac{N-1}{N}\right)^2 \left(\|x_i - \bar{x}_{-i}\|^2 - \|x_i' - \bar{x}_{-i}\|^2\right) \\ = \|x_i - \bar{x}\|^2 - \|x_i' - \bar{x}'\|^2. \quad (4)\end{aligned}$$

Now, returning to the second part of the change in the potential function, we can use the formal definition of $n_j(X)$ to write:

$$\sum_{j=1}^{N} \frac{1}{2}\lambda \left(n_j(X) - n_j(X')\right) = \frac{1}{2}\lambda \frac{1}{m} \sum_{\mu} \sum_{j=1}^{N} \sum_{k \neq j} \left(\delta(x_{j,\mu}, x_{k,\mu}) - \delta(x'_{j,\mu}, x'_{k,\mu})\right).$$

The terms cancel whenever $j \neq i$ and $k \neq i$, so we are left with:

$$\begin{aligned}\sum_{j=1}^{N} \frac{1}{2}\lambda \left(n_j(X) - n_j(X')\right) = \\ \frac{1}{2}\lambda \frac{1}{m} \sum_{\mu} \left(\sum_{j \neq i} \left(\delta(x_{j,\mu}, x_{i,\mu}) - \delta(x'_{j,\mu}, x'_{i,\mu})\right) + \sum_{k \neq i} \left(\delta(x_{i,\mu}, x_{k,\mu}) - \delta(x'_{i,\mu}, x'_{k,\mu})\right) \right) \\ = \lambda \frac{1}{m} \sum_{\mu} \sum_{j \neq i} \left(\delta(x_{j,\mu}, x_{i,\mu}) - \delta(x'_{j,\mu}, x'_{i,\mu})\right) \\ = \lambda \frac{1}{m} \sum_{\mu} \left(n_{i,\mu}(X) - n_{i,\mu}(X')\right) \; = \; \lambda \left(n_i(X) - n_i(X')\right). \quad (5)\end{aligned}$$

Putting Equations (4) and (5) together, we have now shown that $\Phi(X') - \Phi(X) = u_i(X') - u_i(X)$. □

Proof of Theorem 1

Theorem 1 now follows from Lemma 1 by Monderer and Shapley's argument (1996b). □

Supplemental Figures

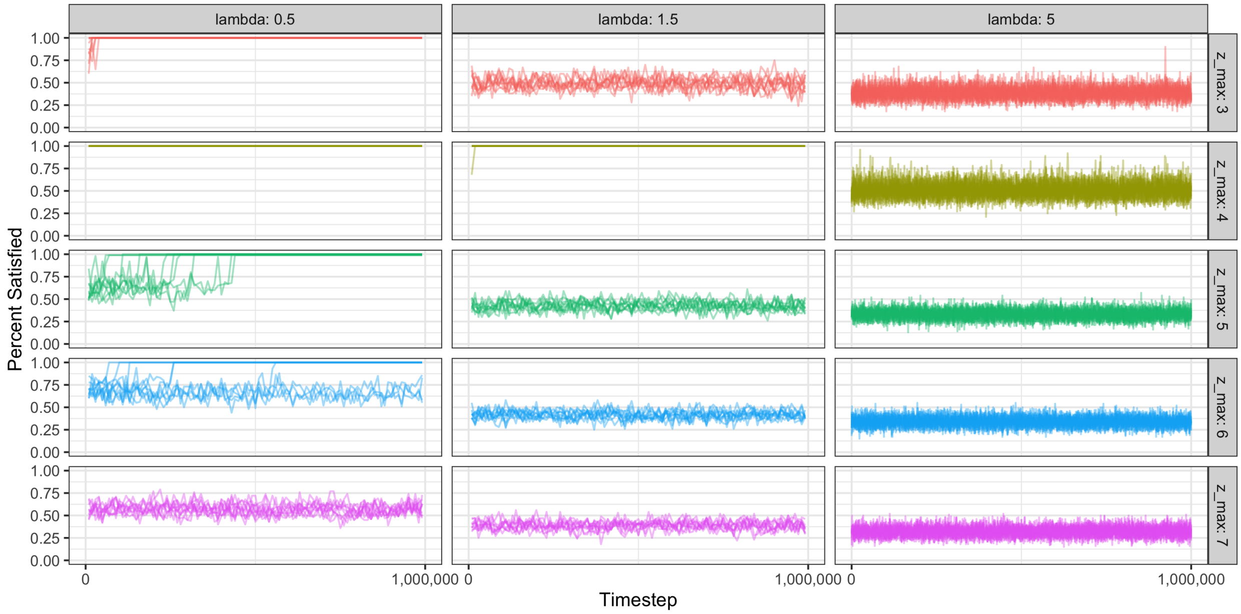

Figure SM1: Percentage of individuals satisfied over 1,000,000 time steps for each trial with $m = 1$, $d = 2$, and varying λ, for networks with $r = 1$ and varying $z_{\max}$.

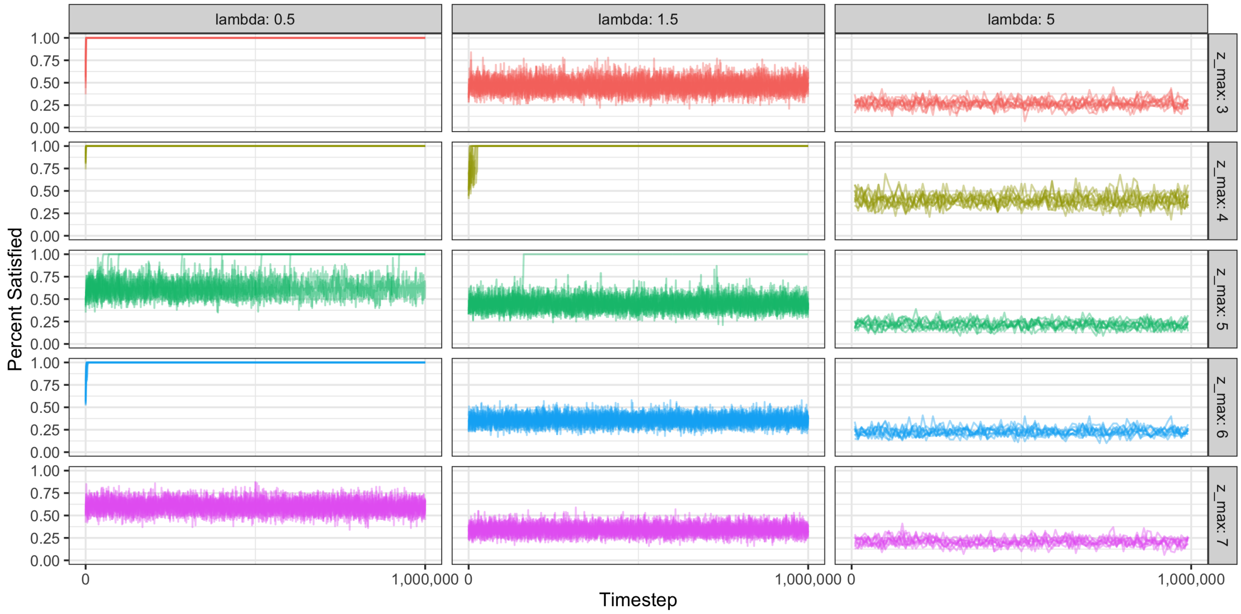

Figure SM2: Percentage of individuals satisfied over 1,000,000 time steps for each trial with $m = 2$, $d = 1$, and varying λ, for networks with $r = 1$ and varying $z_{\max}$.

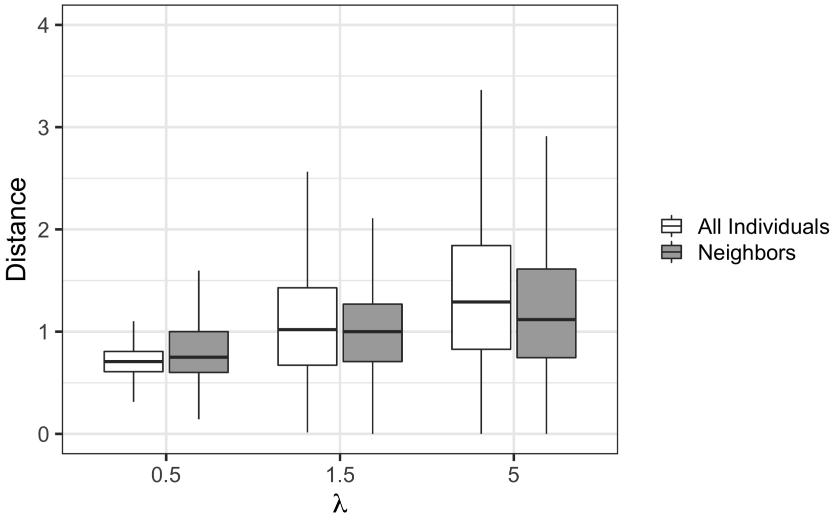

Figure SM3: Box plots showing distances from individuals' identities to the average identity of all individuals in the population and to the average identity of their neighbors in the network, measured at the $10,000$th time step, for $m = 1$, $d = 2$, and varying λ, aggregating trials across the different networks. The differences between the average distance to the population mean identity and the average distance to the mean of one's neighbors' identities are all significant with $p < .001$ in paired t-tests: $t(249799) = -50.21$ when $\lambda = 0.5$; $t(249799) = 82.67$ when $\lambda = 1.5$; and $t(249799) = 123.33$ when $\lambda = 5$.

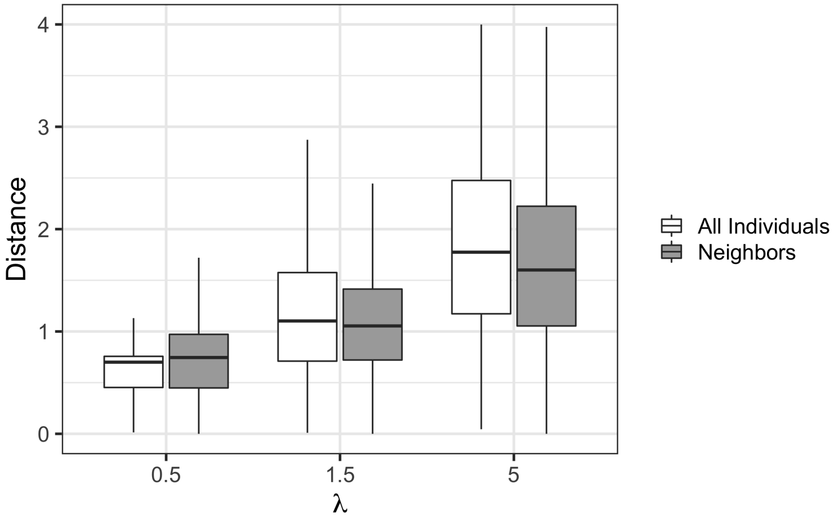

Figure SM4: Box plots showing distances from individuals' identities to the average identity of all individuals in the population and to the average identity of their neighbors in the network, measured at the $10,000$th time step, for $m = 2$, $d = 1$, and varying λ, aggregating trials across the different networks. The differences between the average distance to the population mean identity and the average distance to the mean of one's neighbors' identities are all significant with $p < .001$ in paired t-tests: $t(2472) = -10.22$ when $\lambda = 0.5$; $t(249799) = 86.50$ when $\lambda = 1.5$; and $t(2497) = 12.10$ when $\lambda = 5$.