Experimental Evidence for Tipping Points in Social Convention

Damon Centola[1,2*], Joshua Becker[1], Devon Brackbill[1] and Andrea Baronchelli[3]

[1]Annenberg School for Communication, University of Pennsylvania; [2] School of Engineering, University of Pennsylvania;
[3] Department of Mathematics, City, University of London, *Correspondence to: dcentola@asc.upenn.edu

Abstract: Theoretical models of critical mass have shown how minority groups can initiate social change dynamics in the emergence of new social conventions. Here we study an artificial system of social conventions in which human subjects interact to establish a new coordination equilibrium. The findings provide direct empirical demonstration of the existence of a tipping point in the dynamics of changing social conventions. When minority groups reached the critical mass –that is, the critical group size for initiating social change –they were consistently able to overturn the established behavior. The size of the required critical mass is expected to vary based on theoretically identifiable features of a social setting. Our results show that the theoretically predicted dynamics of critical mass do in fact emerge as expected within an empirical system of social coordination.

Observational accounts of rapid changes in social conventions have suggested that apparently stable societal norms can be effectively overturned by the efforts of small but committed minorities (*1–3*). From social expectations about gender roles in the workplace (*4*), to the popular acceptance of (or intolerance toward) tobacco use and marijuana use (*5*), accounts of changing social conventions have hypothesized that minority groups can trigger a shift in the conventions held by the majority of the population (*1–3*, *5*, *6*). While this hypothesis presents a striking contrast to the expectations of classical equilibrium stability analysis from economic theory (7,8), it can nevertheless be well-explained by the theory of critical mass as posited by evolutionary game theory (*9–11*). This theory argues that when a committed minority reaches a critical group size – commonly referred to as a "critical mass" – the social system crosses a tipping point. Once the tipping point is reached, the actions of a minority group trigger a cascade of behavior change that rapidly increases the acceptance of a minority view (*12–14*).

The simplest formulation of critical mass theory maintains that small groups of regular individuals –that is, with the same amount of social power and resources as everyone else –can successfully initiate a change in social conventions. According to this view, the power of small groups comes not from their authority or wealth but from their commitment to the cause (*14*, *15*).

Thus far, evidence for critical mass dynamics in changing social conventions has been limited to formal theoretical models, and observations from qualitative studies. These studies have proposed a wide range of possible thresholds for the size of an effective critical mass, ranging from 10% of the population up to 40%. For instance, theoretical simulations of linguistic conventions have argued that a critical mass composed of 10% of the population is sufficient to overturn an established social equilibrium (*14*). By contrast, qualitative studies of gender conventions in

corporate leadership roles have hypothesized that tipping points are only likely to emerge when a critical mass of 30% of the population is reached (*3*, *16*). Related observational work on gender conventions (*17*) has built on this line of research, speculating that effective critical mass sizes are likely to be even higher, approaching 40% of the population. Despite the broad practical (*18*, *19*) and scientific (*1*, *12*) importance of understanding the dynamics of critical mass in collective behavior, it has not been possible to identify whether there are in fact tipping points in empirical systems because such a test requires the ability to independently vary the size of minority groups within an evolving system of social coordination.

We addressed this problem by adopting an experimental approach to studying tipping point dynamics within an artificially created system of evolving social conventions. Following the literature on social conventions (*9*, *20*, *21*), we study a system of coordination in which a minority group of actors attempt to disrupt an established equilibrium behavior. In both our theoretical framework and empirical setting, we adopt the canonical approach of using coordination on a naming convention as a general model for conventional behavior (*21–24*). Our experimental approach is designed to test a broad range of theoretical predictions derived from the existing literature on critical mass dynamics in social conventions.

We first synthesized these diverse theoretical and observational accounts of tipping point dynamics to derive theoretical predictions for the size of an effective critical mass (*25*). Based on earlier theoretical (*9*, *26*) and qualitative studies of social convention (*20*, *23*), we propose a simple model of strategic choice in which actors decide which social conventions to follow by choosing the option that yields the greatest expected individual reward given their history of social interactions (*9*). In this individual learning model, people coordinate with their peers so long as they benefit individually from coordinating. The model predicts a sharp transition in the collective dynamics of social convention as the size of the committed minority reaches a critical fraction of the population (Fig. 1). When the size of the committed minority is below this predicted tipping point, the dominant social convention is expected to remain stable, while above this size it is expected to change (*25*).

Our theoretical predictions for the size of the critical mass were determined by two parameters: individual memory length (M), and population size (N). Explorations of these parameters (Fig. 1) show that the predicted size of the tipping point changes significantly with individuals' expected memory length (M). When participants have shorter memories (M <5 interactions), the size of the critical mass is smaller. Even under the assumption that people have very long memories (M >100 interactions), the predicted critical mass size remains well below 50% of the population (*25*), indicating that critical mass dynamics may be possible even in systems with long histories. Variations in population size were explored computationally in the range $20<N<100{,}000$ and were not found to significantly affect the predicted critical mass size (*25*). Figure 1 shows that for populations in the range $20<N<1000$ stochastic fluctuations introduce a small uncertainty into the estimate of the critical mass size. However, for population sizes $N>1000$, the predicted tipping point for social change is constant and independent of N (complete details in (*25*)).

We recruited 194 subjects from the World Wide Web, and placed them into online communities where they participated in a social coordination process (*27*, *28*). Upon arrival to the study, participants were randomly assigned to participate in one of 10 independent online groups, which varied in size from 20 to 30 people. In a given round of the study, two members of each group were chosen at random to interact with one another. Both subjects simultaneously assigned names to a pictured object (i.e., a face), attempting to coordinate in the real-time exchange of linguistic alternatives (*20*, *25*). If the players entered the same name (i.e., coordinated), they were rewarded with a successful payment; if they entered different names (i.e., failed to coordinate), they were penalized. In each community, individuals interacted with each other over repeated rounds of randomly assigned pairings, with the goal of coordinating with one another (*25*). Participants were not incentivized to reach a "global" consensus, but only to coordinate in a pairwise fashion with their partner on each round. Participants were financially rewarded for coordinating, and financially punished each time they failed to coordinate with each other (*25*). Once a convention was established for the entire population, the incentives strongly favored coordinating on the equilibrium behavior.

After each round, the participants could see only the choices that they and their partner had made, and their cumulative pay was updated accordingly. They were then randomly assigned to interact with a new member of their group, and a new round would begin. These dynamics reflect common types of online exchanges, in which community members directly interact the other members of a large, often anonymous population –using, for instance, chat interfaces or messaging technologies –leading them to adopt linguistic and behavioral conventions that allow them to effectively coordinate their actions with other participants' expectations (*20*, *29*, *30*). Consistent with these types of settings, participants in the study did not have any information about the size of the population that was attempting to coordinate nor about the number of individuals to whom they were connected (*9*, *20*, *23*). In every group, this interaction process quickly led to the establishment of a group-wide social convention, in which all players in the network consistently coordinated on the same naming behavior (*20*, *25*). Once a convention was established among all experimental participants, we introduced a small number of confederates (that is, a "committed minority") into each group, who attempted to overturn the established convention by advancing a novel alternative (*25*).

Trials varied according to the size of the committed minority (C) that attempted to overturn the established convention. In total, we studied the dynamics of critical mass in 10 independent groups, each with a committed minority of a fixed size. Across all 10 groups, the sizes of the committed minorities were in the range ($15\% < C < 35\%$).

Figures 2 and 3 report tipping point dynamics in the collective process of overturning an established equilibrium. Consistent with the expectations of our theoretical model (using empirically parameterized values of N and M), when the size of the committed minority reached approximately 25% of the population, a tipping point was triggered, and the minority group succeeded in changing the established social convention.

Five trials were conducted. Each trial was composed of two communities –one with the committed minority below the expected critical size ($C < 25\%$), and one with it above ($C > 25\%$). In every trial, the community with $C < 25\%$ had only small numbers of converts to the minority view. Over the course of these trials, each of these converts eventually reverted back to the dominant norm. Continuous interactions led to occasional switching by subjects throughout the study. However, over all of the trials, in the condition where the minority group was smaller than 25% of the population, on average only 6% of the non-committed population adopted the alternative behavior by the final round of the study.

For each of these unsuccessful trials, we conducted a corresponding trial using another population of the same size, but with a larger committed minority ($25\% \leq C \leq 31\%$). In all of these groups, the alternative norm reached the majority of the population within the experimental window of observation (Figures 2 and 3). Over all trials, populations with $C \geq 25\%$ were significantly more likely to overturn the dominant convention than populations with a committed minority below 25% (P=.01, Wilcoxon rank sum). We found that in one case (Trial 1) this transition from failure to success was the result of increasing the size of the committed minority by only one person.

Figure 3 shows a summary of final adoption levels across all trials, along with expectations from our empirically parameterized theoretical model, with 95% confidence intervals. Populations with committed minorities ranging from $25\% \leq C \leq 27\%$ achieved uptake levels between 72% and 100% within the empirical observation window. At C=31%, the committed minority achieved consensus within the window of empirical observation. Figure 3 compares these observations to numerical simulations of the theoretical model using population sizes and observation windows comparable to the experimental study (N=24, T=100, M=12). Memory length for these simulations was calibrated using subjects' empirical memory lengths in this study based on their observed behavior over all 10 groups. A memory length in the range $9 \leq M \leq 13$ provides a good approximation of subjects' observed behavior, correctly predicting 80% of subjects' choices across all trials (*25*). The theoretically predicted critical mass size from this model fit the experimental findings well (Fig. 3). Numerical analyses indicate that with larger population sizes the critical mass point becomes more exact (See Fig. 1 and Fig. 3), approaching 24.3% of the population.

Our experimental results do not show agreement with theoretical predictions from models of social convention that predict low critical mass thresholds, at 10% of the population. However, our findings show good agreement with qualitative studies of gender conventions within organizational settings (*3*), which hypothesized that a critical mass of approximately 30% could be sufficient to overturn established norms (*16*). Our results may suggest that in organizational contexts –where population boundaries are relatively well-defined, and there are clear expectations and rewards for social coordination among peers –the process of normative changes in social conventions may be well-described by the dynamics of critical mass.

The design choices that aided our control of the study also put constraints on the behaviors that we could test. Our experimental design provided subjects with social and financial incentives that strongly favored coordinating on an established social convention (*25*). However, in the real

world, individuals' emotional and psychological commitments to established behaviors can create additional resistance to behavior change (*31*). To further explore these expectations, supplementary analyses of our theoretical model (Fig. S7) (*25*) extend our basic predictions to consider how the critical mass size may differ under conditions of greater social entrenchment. When actors are more conservative –exhibiting an explicit bias in favor of the established convention (based on a skewed best response calculation favoring the equilibrium behavior) – tipping point dynamics were still predicted to be achievable by committed minorities with only marginally larger group sizes.

In delimiting the scope of our findings, we emphasize that the critical mass value of 25% is not expected to be a universal value for changing social conventions. Our results demonstrate that within an endogenous system of social coordination, tipping point dynamics emerged consistent with theoretical expectations. Further work is required to determine the applicability of our findings to specific social settings. In particular, alternative empirical parameterizations of our model can result in alternative predictions for the expected size of the critical mass. We expect that the findings from our study can be significantly expanded by future empirical work studying the dynamics of tipping points within other empirical systems of social convention.

For instance, an important setting in which these results might be usefully applied concerns the growing ability of organizations and governments to use confederate actors within online spaces to influence conventional behaviors and beliefs. Recent work on the 50c party in China (*32*, *33*) has argued that the Chinese government has incentivized small groups of motivated individuals to anonymously infiltrate social media communities such as Weibo with the intention of subtly shifting the tone of the collective dialogue to focus on topics that celebrate national pride, and distract from collective grievances (*32*). We anticipate that social media spaces of this kind will be an increasingly important setting for extending the findings of our study to understand the role of committed minorities in shifting social conventions. Similarly, the results from our study may also be usefully applied to the dynamics of critical mass in other online settings, such as changing social expectations regarding *i*) the standards of civility in Facebook and other online discussion forums (*19*, *34*), *ii*) the acceptability of bullying behavior in adolescent chat groups (*35*), and *iii)* the appropriate kinds of content to share with strangers over social media (*36*), all of which have been suggested to exhibit susceptibility to shifts in conventional behavior as a result of the activity of a small fraction of the population (*19*, *34*, *36*).

References and Notes:

1. T. Kuran, The inevitability of future revolutionary surprises. *Am. J. Sociol.* **100**, 1528–1551 (1995).
2. K. D. Opp, C. Gern, Dissident groups, personal networks, and spontaneous cooperation: The East German revolution of 1989. *Am. Sociol. Rev.*, 659–680 (1993).
3. R. M. Kanter, Some effects of proportions on group life: Skewed sex ratios and responses to token women. *Am. J. Sociol.* **82**, 965–990 (1977).
4. D. Dahlerup, L. Freidenvall, Quotas as a 'fast track'to equal representation for women: Why Scandinavia is no longer the model. *Int. Fem. J. Polit.* **7**, 26–48 (2005).
5. S. Bikhchandani, D. Hirshleifer, I. Welch, A theory of fads, fashion, custom, and cultural change as informational cascades. *J. Polit. Econ.*, 992–1026 (1992).
6. T. Kuran, *Private truths, public lies: The social consequences of preference falsification* (Harvard University Press, 1997).
7. J. C. Harsanyi, R. Selten, *A general theory of equilibrium selection in games* (MIT Press, Cambridge, 1988).
8. J. F. Nash, Equilibrium points in n-person games. *Proc. Natl. Acad. Sci.* **36**, 48–49 (1950).
9. H. P. Young, The evolution of conventions. *Econom. J. Econom. Soc.*, 57–84 (1993).
10. M. Kandori, G. J. Mailath, R. Rob, Learning, mutation, and long run equilibria in games. *Econom. J. Econom. Soc.*, 29–56 (1993).
11. G. Ellison, Learning, local interaction, and coordination. *Econom. J. Econom. Soc.*, 1047–1071 (1993).
12. T. Schelling, *Micromotives and macrobehavior* (WW Norton & Company, New York, 1978).
13. M. Granovetter, Threshold models of collective behavior. *Am. J. Sociol.*, 1420–1443 (1978).
14. J. Xie *et al.*, Social consensus through the influence of committed minorities. *Phys. Rev. E.* **84**, 011130 (2011).
15. D. M. Centola, Homophily, networks, and critical mass: Solving the start-up problem in large group collective action. *Ration. Soc.* **25**, 3–40 (2013).
16. D. Dahlerup, From a small to a large minority: women in Scandinavian politics. *Scand. Polit. Stud.* **11**, 275–298 (1988).
17. S. Grey, Numbers and beyond: The relevance of critical mass in gender research. *Polit. Gend.* **2**, 492–502 (2006).
18. G. Marwell, P. Oliver, *The critical mass in collective action* (Cambridge University Press, 1993).
19. K. Nyborg *et al.*, Social norms as solutions. *Science*. **354**, 42–43 (2016).
20. D. Centola, A. Baronchelli, The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proc. Natl. Acad. Sci.* **112**, 1989–1994 (2015).
21. D. Lewis, *Convention: A philosophical study* (Harvard University Press, Cambridge, 1969).
22. C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591 (2009).

23. S. Garrod, G. Doherty, Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*. **53**, 181–215 (1994).
24. L. Wittgenstein, *Philosophical investigations* (John Wiley & Sons, 2009).
25. Materials and methods are available as supporting materials with this submission.
26. A. Baronchelli, M. Felici, V. Loreto, E. Caglioti, L. Steels, Sharp transition towards shared vocabularies in multi-agent systems. *J. Stat. Mech. Theory Exp.* **2006**, P06014 (2006).
27. D. Centola, An experimental study of homophily in the adoption of health behavior. *Science*. **334**, 1269–1272 (2011).
28. D. Centola, The spread of behavior in an online social network experiment. *science*. **329**, 1194–1197 (2010).
29. L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan, Group formation in large social networks: membership, growth, and evolution. *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 44–54 (2006).
30. F. Kooti, H. Yang, M. Cha, P. K. Gummadi, W. A. Mason, The Emergence of Conventions in Online Social Networks. *Proc. Sixth Int. AAAI Conf. Weblogs Soc. Media ICWSM 2012* (2012).
31. C. Tilly, S. Tarrow, *Contentious Politics* (Oxford University Press, New York, ed. 2nd, 2015).
32. G. King, J. Pan, M. E. Roberts, How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *Am. Polit. Sci. Rev.* **111**, 484–501 (2017).
33. G. King, J. Pan, M. E. Roberts, Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*. **345**, 1251722 (2014).
34. A. Antoci, A. Delfino, F. Paglieri, F. Panebianco, F. Sabatini, Civility vs. Incivility in Online Social Interactions: An Evolutionary Approach. *PLOS ONE*. **11**, e0164286 (2016).
35. E. L. Paluck, H. Shepherd, P. M. Aronow, Changing climates of conflict: A social network experiment in 56 schools. *Proc. Natl. Acad. Sci.* **113**, 566–571 (2016).
36. C. Shih, *The Facebook era: Tapping online social networks to market, sell, and innovate* (Pearson Education, 2010).
37. S. Maslov, K. Sneppen, U. Alon, S. Bornholdt, H. G. Schuster, *Handbook of graphs and networks* (Wiley-VCH New York:, 2003).

Acknowledgements: The authors gratefully acknowledge research assistance from Devon Brackbill, Soojong Kim, and Natalie Herbert, and programming assistance from Alan Wagner and Ryan Overbey.

Institutional Review Board: This research was approved by the Institutional Review Board at the University of Pennsylvania, where the study was conducted, and it included informed consent by all participants in the study.

Data Availability: The data for this study are available at : http://ndg.asc.upenn.edu/wp-content/uploads/2018/03/Centola-Becker-Baronchelli_Complete-Data.pdf

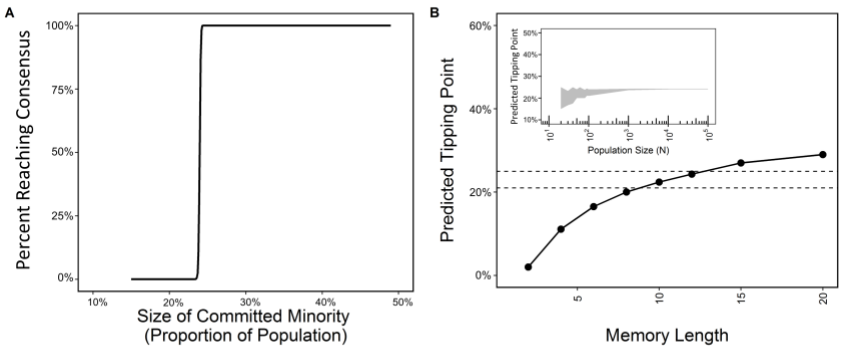

Fig. 1. Predicted tipping points in social stability. A) Theoretical modeling of the proportion of outcomes in which the alternative behavior is adopted by 100% of the population. In this system, the number of agents **(*N*)=1000,** the number of interactions **(*T*)=1000,** the number of past interactions used in agent decisions ***(M)*=12)**. **B)** The size of the predicted critical mass point is shown as a function of individuals' average memory length, *M*, where **(*N*=1000, *T*=1000)**. The dashed lines indicate the range enclosed by our experimental trials, showing the largest unsuccessful minority (21%) and the smallest successful minority (25%). Although the expected size of the critical mass point increases with *M*, this relationship is concave, allowing the predicted tipping point to remain well below 50% as *M* gets large (>100). **Inset** Effect of increasing population size on the precision of the size of the committed minority (*C*) prediction **(*M*=12, *T*=1000)**. For *N*<1000, small variations in the predicted tipping point emerge due to stochastic variations in individual behavior. Shaded region indicates *C* sizes where success was observed frequently, but without certainty. Above this region, for larger *C* sizes, the probability of success reaches 1; for *C* sizes below this region, the likelihood of success goes to 0.

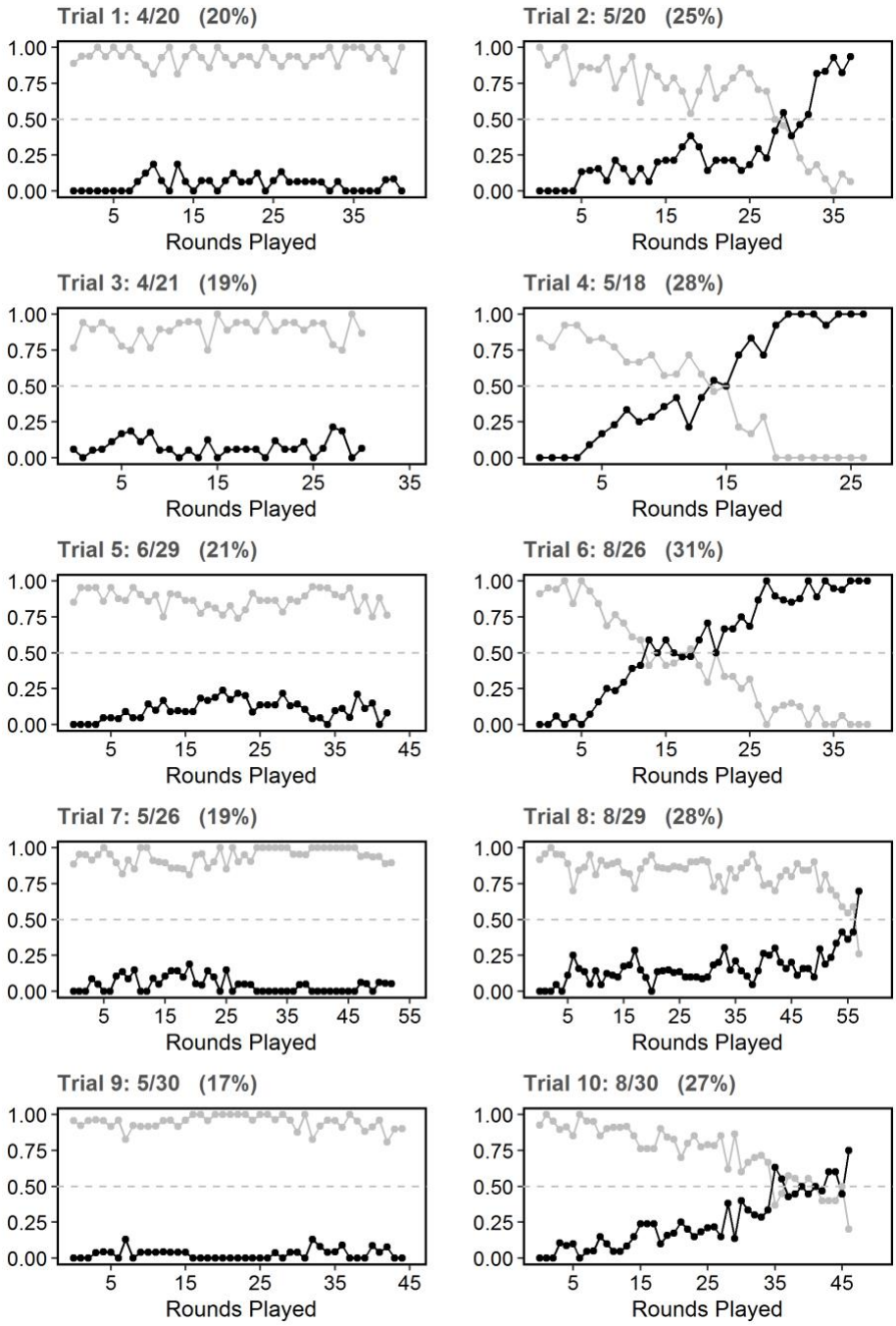

Fig. 2. Time series showing adoption of the alternative convention by non-committed subjects (i.e., experimental subjects). Gray line indicates the popularity of the established convention and black line shows the adoption of the alternative convention. Success was achieved when more than 50% of the non-committed population adopted the new social convention. Trials in the left column show failed mobilization, while trials in the right column show successful mobilization. A transition in the collective dynamics happens when *C* reaches approximately 25% of the population. Each round is measured as N/2 pairwise interactions, such that each player has one interaction per round on average.

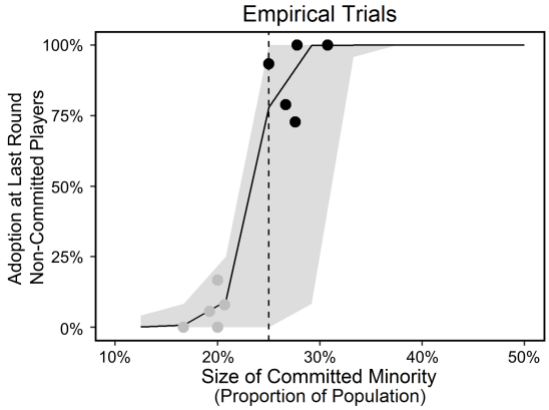

Fig. 3. Final success levels from all trials (gray points indicate trials with $C < 25\%$; black points indicate trials with $C \geq 25\%$). Also shown is the theoretically predicted critical mass point (solid line) with 95% confidence intervals (***N*=24, *T*=45, *M*=12**; gray area indicates 95% confidence intervals from 1,000 replications). The dotted line indicates $C=25\%$. The theoretical model of critical mass provides a good approximation of the empirical findings. For short time periods ($T<100$), the critical mass point prediction is not exact (ranging from $20\% < C < 30\%$ of the population); however over longer time periods ($T > 1000$) the transition dynamics become more precise (solid line, SOM).

Supplementary Materials

Materials and Methods

Experimental Design

Subjects were randomly assigned to a fully-connected network (i.e., homogeneously mixing population) containing a pre-determined number of experimental participants and confederates. The number of confederates as a fraction of the total population was varied between conditions. Our theoretical model predicts a tipping point with approximately 25% of the population. To test this effect, we ran 5 trials with fewer than 25% confederates ("below threshold") and five trials with greater than 25% confederates ("above threshold"). Because users had no way of distinguishing confederates from experimental participants, the user experience in below-threshold conditions was identical to the user experience in above-threshold conditions.

Experimental trials consisted of two phases. Once a trial began, the first phase of the experimental procedure (see "Subject Experience During the Experiment" below) allowed subjects to interact until they established a shared social convention. This phase used the procedures from previous research on the emergence of conventions *(20)*. During this phase, confederate participants entered no response, except during trials 3 and 4, in which they played a single pre-determined strategy during phase one to speed up initial convergence. Phase one was considered complete when every experimental participant began using the same strategy, with at most two players deviating. This determination was based on previous experimental methods for identifying convergence on social conventions *(20)*, which indicated that even once a stable convention is established, not all players will always use the conventional strategy.

Once an endogenous convention was reached, the confederate participants simultaneously began entering a pre-determined response that differed from the established convention. All the confederates used the same response. In order to ensure that the alternative response was comparable to the established convention, we used responses selected from names that commonly emerged as conventions in previous studies (e.g., "Mary"). Confederate participants continued entering this alternative strategy until the game was complete.

The length of an entire experimental trial was determined prior to play beginning. Due to stochastic variation in the length of the first experimental phase, consistent with previous research on the emergence of conventions *(20)*, the number of interactions available for Phase 2 varied between trials. The number of interactions also varied because subjects entered responses at different rates, and the trials were run until each player participated in a minimum number of interactions.

Subject Recruitment

Participants in our study were recruited via the World Wide Web to be players in online games for research purposes. Players registered by providing their email address and selecting a username and an avatar. All players were required to provide informed consent in order to register. Once registered, users were placed into a recruitment pool for future experiments. Registered users were then sent an advertisement and a link to participate in a trial for this study. Trials were run over a 105 day period between July 15, 2015, and October 7, 2015.

Subject Experience During the Experiment

Upon arriving at the study website, participants viewed instructions on how to play the game (Figure S1). In the game, participants play a series of one-shot coordination games, as shown in Figure S2. The left hand column shows the other players in the game. However, they have no information about which player ("partner") they are matched with for a given interaction. Participants cannot identify whether they are matched with another experimental participant or a confederate. They cannot identify if their partner for a given interaction is the same as their partner in a previous interaction.

For each one-shot game, participants are shown an image (a picture of a man or a woman) and prompted to enter a name in the response field. There is one picture per trial—i.e., the picture does not change across interactions and remains the same for all the interactions within the same experimental trial. Participants are instructed that if they enter the same name as their partner, they will be rewarded $0.10. If they do not enter the same name as their partner, $0.10 is deducted from their total winnings, with a minimum possible reward of $0.00. On the right hand side, participants are shown their progress in the game. Completed interactions indicate whether the interaction was a match. After each interaction, participants are shown the name entered by their partner, regardless of whether or not it was a match. Centola & Baronchelli *(20)* showed that by this process, subjects will successfully establish a shared convention; i.e., this process leads subjects to converge on a single shared name for the person depicted in the image.

Supplementary Text

Model Definition

Our theoretical model follows previous theoretical models of critical mass in studying asynchronous pairwise interaction *(14,26)*. However, while these prior models assume that agents randomly select from previously observed strategies, we follow game theoretic models of convention *(9)* in modeling strategic choice in which individuals attempt to choose the behavior most likely to generate successful coordination.

In each time step, two agents from a population of N agents are randomly selected to interact. One agent is randomly selected to be the "speaker" and the other agent is assigned the role of "hearer." The agent playing the role of speaker picks a best response strategy. The best response strategy is defined as the strategy most frequently observed in previous interactions in which that agent was the hearer. An agent's "memory" stores a record of the strategies observed in use by other players, and an agent only updates their memory during interactions in which they are the hearer. Agents do not respond to a complete history of past plays; rather, we assume that agents determine their best response strategy based only on the past M interactions. The agent decision rule is therefore defined by the single parameter M that determines the size of an agent's memory. This limit reflects both the assumption that agents have limited cognitive resources and also the assumption that recent interactions are more informative of population behavior *(9)*.

The formalization of memory as a sliding window within a vector of past plays was chosen in order both to be consistent with previous theoretical research in evolutionary game theory *(9)* and also to maintain a parsimonious model with minimal degrees of freedom. Other possible formalizations for memory include a continuous decay model. This model was not selected for this study because it would involve more degrees of freedom to specify a decay function and associated parameters. As a simple approximation, we find that a sliding window formalization

correctly predicts approximately 80% of user plays. Thus far, we have not found that any alternative model provides a better fit with the empirical data.

Following a standard procedure from the literature *(14)*, we model a scenario in which a population has already converged on some convention, so that every agent uses some previously established strategy 'A', such that A is the only option contained in each agent's memory. The simulation is therefore initialized so that every non-committed agent has a memory that contains only A (i.e., each non-committed agent begins the simulation with a memory as if they had observed A for the previous M interactions). Our simulation studies the dynamics that occur when some fraction of that population begins to use an alternative strategy 'B' instead of following best-response dynamics. This committed minority always uses a single fixed strategy B when they are selected as speaker instead of using a best response strategy. Thus, we simulate the effect of a committed minority using strategy B in a population with an established convention A. With the exception of the decision heuristic used by non-committed agents, our model is identical to that studied by Xie et al. *(14)*, which is itself identical to Baronchelli et al. *(26)*. The best-response decision-rule for non-committed agents is identical to that studied by Young *(9)*.

We consider a population of N agents. Some fraction C of the total N agents are identified as committed agents who always play B. At time T=0, the agents playing best-response dynamics (i.e., non-committed agents) are initialized with a memory vector of length M, the entries of which are all A. The model is fully defined by the following parameters:

- N: the number of agents
- C: the fraction of the population belonging to the committed group
- M: the number of past interactions used in agent decisions

This model defines a Markov chain with only one absorbing state: the state in which the entire population has adopted the strategy promoted by the committed minority. However, this will only occur in infinite time when the committed minority is small *(14)*. In finite time, the model is characterized by two states. Above the tipping point, the alternative strategy promoted by the committed minority is very quickly adopted by the entire population. Below the tipping point, the model reaches a quasi-stationary state in which the initial convention is the dominant convention for very long periods of time *(14)*.

We therefore measure the tipping point in simulation by simulating T=1000*N interactions (i.e., to allow an average of 1000 interactions per agent) and then calculating the percentage of non-committed agents who have adopted the alternative strategy. Results are only minimally perturbed by larger values of T.

Estimating Memory Length

To develop our experimental hypotheses, we used data from previous experiments on the same web platform *(20)* to estimate the value for M by determining the value for which the model most accurately predicts user behavior. We then replicated this analysis for experimental data from this study, producing comparable results. Figure S3 shows the fraction of user choices which are accurately predicted by the model as a function of M. To generate this figure, each user's play was predicted for each interaction based on their M previous observations. We then plot the total fraction of interactions which were correctly predicted for each value of M. Using this parameter estimation, we predict a lower bound of M=12 for users in our experimental platform. Our analyses indicate that there is no reason to think that the participants in our study had longer/shorter memories than anyone else in the general population. To test this, we compared the best-fit memory length estimate for the subjects from earlier studies of naming conventions to the best-fit memory length estimate for the subjects in this study, and we found no significant differences between subject behavior in past studies and subject behavior in the present study.

Tipping Point

Coordination dynamics show a tipping point similar to that observed by Xie et al. *(14)* even after accounting for strategic choice. Using the lower bound estimate M=12, we model short term and long term success of committed groups using strategy B in a population playing A, across a range of committed minority sizes C and population sizes N. Figure S4 shows adoption by non-committed agents after T=45, T=100, and T=1000 "rounds" in a population of N=1,000 agents. To normalize time across variation in population size, each time step ("round") is measured as N interactions, or one interaction per agent. To generate this figure, we initialize the simulation as described above and run for N*T interactions (i.e., T rounds). We measure adoption as the usage of each strategy in the N interactions prior to measurement. We then plot the fraction of non-committed players adopting the alternative strategy after T interactions for each value of C.

Long term adoption at T=1000 rounds shows a sharp transition when committed groups reach approximately 24.2% of the population. Below this threshold, only a small number of non-committed agents use strategy B in the N interactions prior to measurement. Above this threshold, the population reaches an absorbing state in which all agents are consistently using strategy B.

For extremely short term dynamics, committed minorities above the tipping point do not always achieve full convergence on the alternative strategy. For N=1,000, for example, after 20 rounds (Fig S4, light blue line), full convergence is not reached even with C=50%. After 45 rounds (the average length of experimental trials), most above-threshold groups are successful, while committed groups between 25% and 30% are not yet guaranteed to reach convergence. After 100 rounds, nearly every above-threshold group has achieved widespread adoption.

Robustness to Population Size

The tipping point is robust to variation in population size. To generate figure 1B inset (main text) we run the same analysis shown in Figure S4, but with varying population size. For each population size and each value of C, we measure the percentage of simulations that reach complete convergence – i.e., the percentage of outcomes in which 100% of the non-committed population has adopted the alternative strategy after 1,000 interactions per person.

The grey area in Figure 1B inset (main text) shows values of C and N for which the committed minority is successful (i.e., they achieve full adoption of the alternative strategy) in greater than 1% of simulations, but in fewer than 99% of simulations. That is, for values of C and

N that fall below this area, the minority group never succeeds in changing the social convention, while for values of C that fall above this area the minority group is successful more than 99% of the time.

For small population sizes, the tipping point has the appearance of non-monotonicity with N due to the fact that not all fractions can be converted into a discrete critical mass size: for example, when N=20, a committed group can comprise either 20% (4/20) or 25% (5/20) of the population. Thus, if the tipping point is between 20% and 25%, then it will take a minimum of 25% (5/20) committed individuals to overturn an established norm. For population sizes from 1000 to 100,000 the tipping point stabilizes at a value of 24.2%.

Effect of Memory Parameter

The existence of a tipping point is robust to variation in agent memory length, and a tipping point in long term adoption appears for all values of M. Figure 1B (main text) and Figure S5 show the tipping point as a function of M. To generate these figures, we identify the lowest value for C in which at least 99% simulations reach convergence (i.e., achieve full adoption of the alternative strategy) after N*1000 interactions. As shown in Fig S3 a group larger than 10% required even if agents choose their strategy only based on the previous 4 interactions. The tipping point remains below 50% even for very large values of M, with a critical mass of only 40% required when M=100.

Robustness to Network Density

Our experimental platform studies critical mass dynamics in a homogeneously mixing (fully connected) network, so that any two agents are equally likely to interact. Dynamics are qualitatively similar with sparse random networks, as shown in figure S6. To generate this figure, we generated random networks in which each node has an equal number of connections *(37)*. In sparse networks, the model is defined identically, but instead of uniformly selecting two agents for interaction, the model randomly selects an edge from the network for interaction. One node is then selected as speaker, and one node is then selected as hearer.

Consistent with findings in *(14)*, the tipping point is slightly lower in sparse networks. To generate Figure S6, we run analyses as shown in Figure 4 at varying network densities. At each point, we determine the smallest value for C at which 100% adoption of the alternative strategy is achieved in at least 99% of simulated outcomes. We hold network size constant, and therefore network density is determined by average degree (number of network neighbors) for each node.

Robustness to Agent Strategy Preference

Our experimental design and theoretical model both reflect coordination dynamics in which two agents must decide between two possible strategies, both of which are equally preferred by every non-committed individual in the population. To model the possibility that agents may prefer one strategy over another strategy, we adopt the standard game theoretic formalization of coordination games in which each strategy is assigned a numeric payoff. In our base model, both strategies would be assigned the same payoff since they are equally preferred.

To choose a strategy based on numeric payoff, agents select the strategy with the greatest "expected payoff," which is calculated as the probability of success multiplied by the numeric payoff of a successful interaction. In our computational model of coordination, probability of success is determined by the percentage of recent interactions (where recent is defined by memory length M) in which a particular strategy was observed. For example, if strategy A provides a

payoff of 1 and was observed in 60% of the previous M interactions, and strategy B provides a payoff of 2 and was observed in 40% of the previous M actions, then the payoff for each strategy is calculated as follows:

payoff of strategy B = 0.6 x 1
= 0.6

payoff of strategy A = 0.4 x 2
= 0.8

Thus, strategy A offers a higher expected payoff despite being less frequently observed and offering a lower probability of success.

We use this model of agent preference to test whether we observe critical mass dynamics in situations where agents have a bias towards either the entrenched convention or the alternative strategy. Figure S7 shows the effect of agent preference on adoption dynamics as a function of the relative preference for the established convention, which is measured as the ratio between the payoff for the established convention and the alternative strategy. This figure indicates that even when agents prefer the established convention twice as much as the alternative strategy (i.e., the payoff ratio between the established convention and the alternative strategy equals 2:1) a committed minority can still establish critical mass and achieve widespread adoption of the alternative strategy.

Data Availability

The complete dataset is publicly available for download from the following URL: http://ndg.asc.upenn.edu/experiments/creating-critical-mass/

This dataset contains a time series for each trial starting at phase 2 as described in Materials and Methods. Each row indicates the percentage of responses which are the established convention, the percentage of responses which are the alternative strategy, and the percentage of responses which are any other strategy.

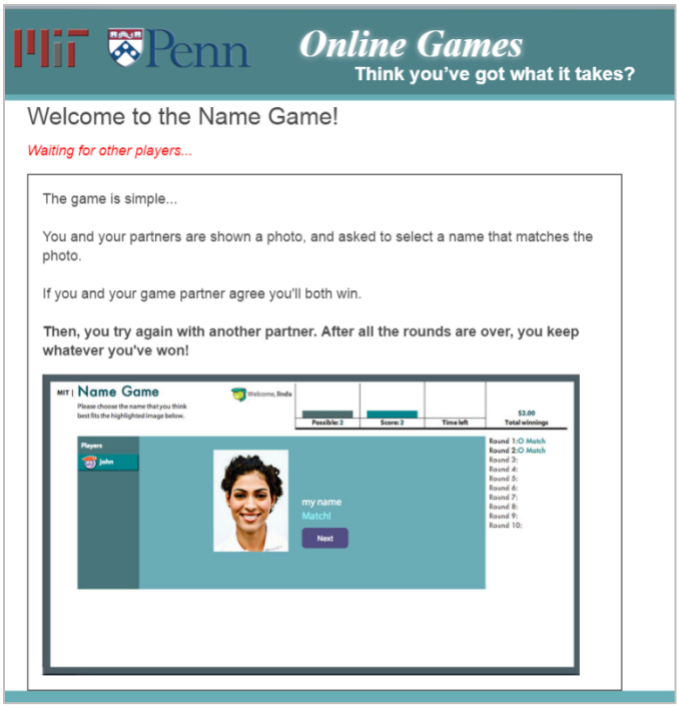

Figure S1. Screenshot of the waiting page showing instructions.

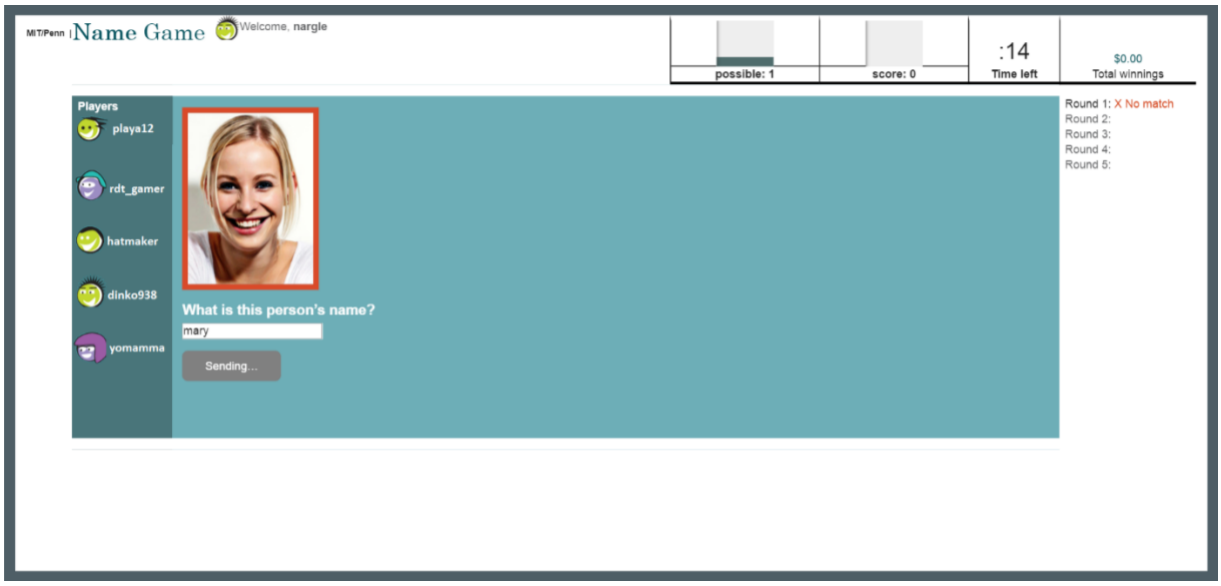

Figure S2. Screenshot of game interface. Note that what is called a "round" in the user interface is a single "interaction" as discussed throughout this text.

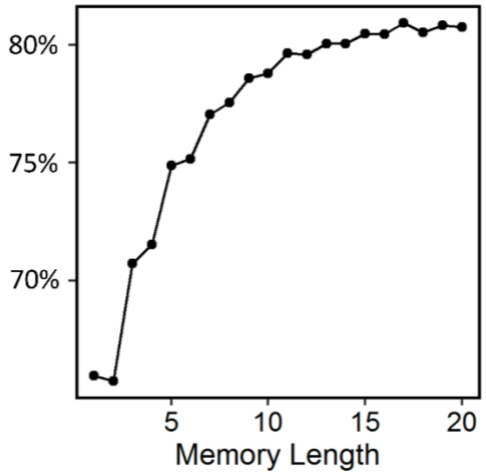

Figure S3. Estimating Memory Length from Empirical Data. Each panel shows the fraction of interactions successfully predicted in the data from the current experiment. When memory length is greater than 10, our model correctly predicts 80% of the choices by our experimental subjects.

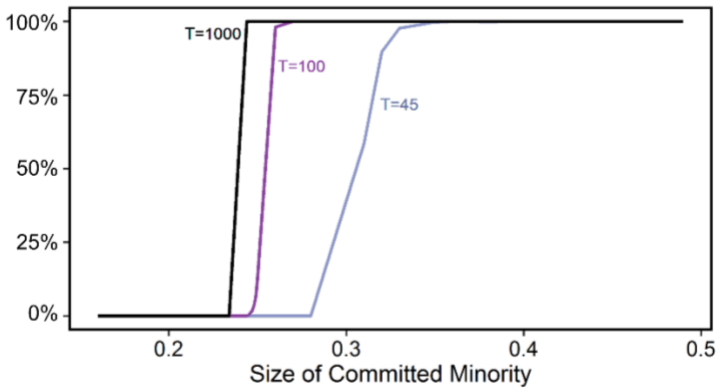

Figure S4. Estimating the Tipping Point. Each line shows the proportion of simulations in which every agent adopted the alternative strategy after T interactions per agent for a network of N=1,000 agents as a function of C. When T=1000, there is a sharp threshold between C=0.241 and C=0.242 indicating that a small change in the size of a committed minority can generate a dramatic shift in the adoption of an alternative convention. It is worth noting that the threshold is well defined only over long time-scales. For shorter time periods, committed minorities that are sufficient in long term dynamics may have not yet achieved convergence.

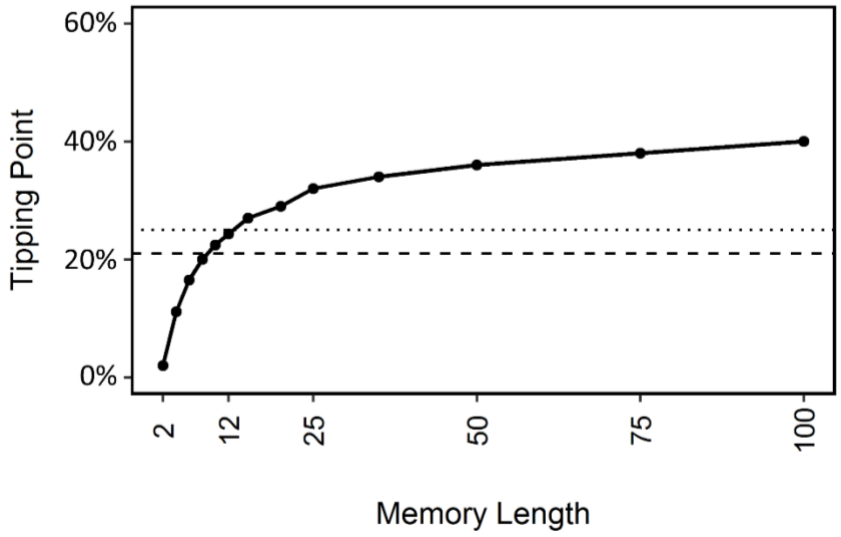

Figure S5. Critical mass dynamics are robust to variation in agent memory length. Each point in this figure shows the tipping point as a function of agent memory length (M). Even when M=100, the tipping point is well below 50%. Horizontal lines indicate the largest value for C which failed in experimental trials (C=21%, dashed line) and the smallest value for C which succeeded in experimental trials (C=25%, dotted line) suggesting that M is between 9% and 13%.

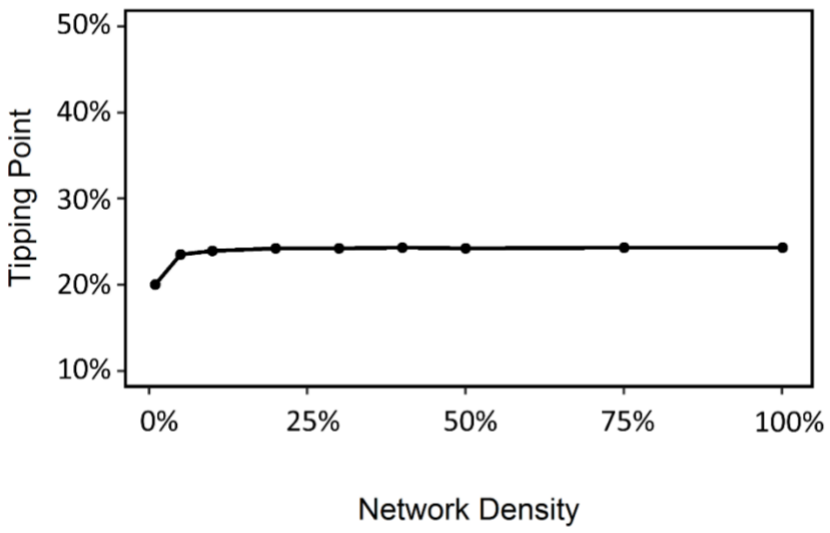

Figure S6. The tipping point is robust to changes in network density. Each point in this figure shows the tipping point as a function of network density for simulations with M=12, T=1000*N, N=1000. In very sparse networks where agents only have a few network neighbors, the tipping point drops slightly but remains above 20%. The point furthest to the left shows simulations for populations where agents have 10 network neighbors, producing a network density of 1%. Network density is defined to be the number of connected edges in a network as a percent of all possible edges that could be connected.

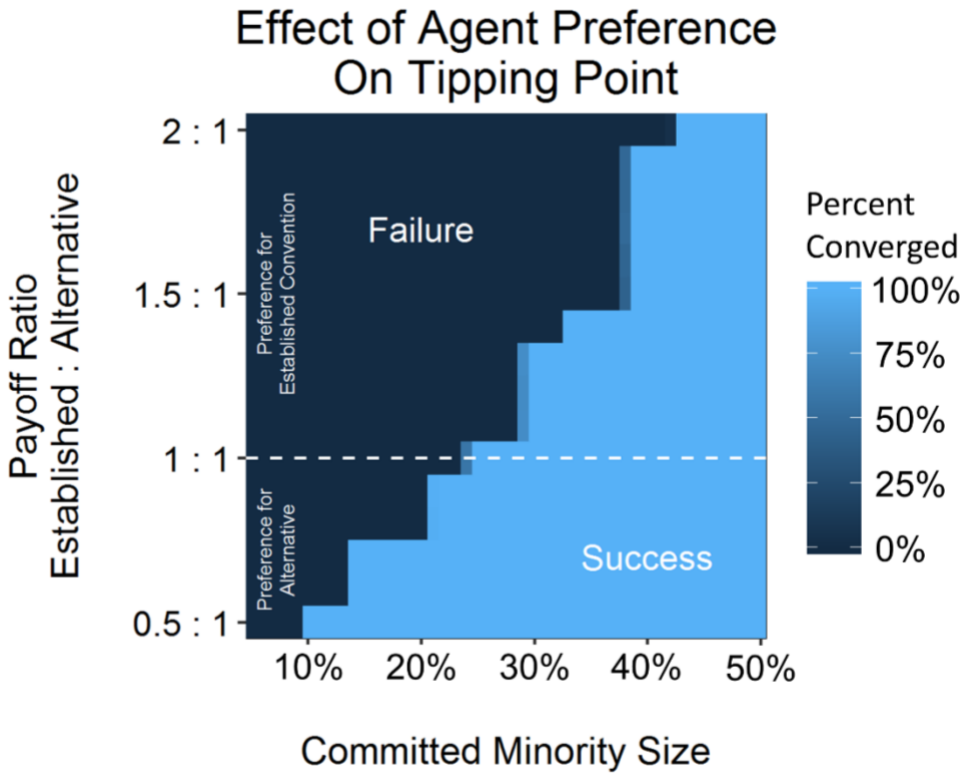

Figure S7. Critical mass threshold as a function of relative preference. This figure shows the proportion of simulations in which the alternative norm is adopted, as a function of minority size and relative preference for each convention for simulations with M=12, N=1000, T=1000*N. The Y axis of this figure indicates the relative payoff of the established convention as compared with the alternative strategy. When the payoff ratio is equal to 1, both strategies are equally desirable (i.e., agents simply adopt the most popularly used strategy) and this model is equivalent to our general model of conventions, showing a tipping point of approximately 25%. When the payoff ratio is equal to 2, agents prefer the established convention twice as much as the alternative strategy, but a committed minority can nonetheless overturn the established convention. When the payoff ratio is equal to 0.5, agents prefer the established convention half as much as the alternative strategy (i.e., they prefer the alternative strategy twice as much as the established convention) and the critical mass can be reached with a very small committed minority comprising less than 10% of the population.