Interdependent Diffusion:
The social contagion of interacting beliefs

Draft October 28, 2020

James P. Houghton
houghton@mit.edu
jameshou@seas.upenn.edu

Abstract:

Social contagion is the process in which people adopt a belief, idea, or practice from a neighbor and pass it along to someone else. For over 100 years, scholars of social contagion have almost exclusively made the same implicit assumption: that only *one* belief, idea, or practice spreads through the population at a time *[1-23]*. It is a default assumption that we don't bother to state, let alone justify. The assumption is so ingrained that our literature doesn't even have a word for "whatever is to be diffused" *[1]*, because we have never needed to discuss more than one of them.

But this assumption is obviously false. Millions of beliefs, ideas, and practices (let's call them "diffusants") spread through social media every day. To assume that diffusants spread one at a time – or more generously, that they spread independently of one another – is to assume that interactions between diffusants have no influence on adoption patterns. This could be true, or it could be wildly off the mark. We've never stopped to find out.

This paper makes a direct comparison between the spread of independent and interdependent beliefs, using simulations and a 2400-subject laboratory experiment. I find that in assuming independence between diffusants, scholars have overlooked social processes that fundamentally change the outcomes of social contagion. Interdependence between beliefs generates polarization, irrespective of social network structure, homophily, demographics, politics, or any other commonly cited cause. It also leads to the emergence of popular worldviews that are unconstrained by ground truth.

These outcomes are at the heart of why we study social contagion: to understand who believes what and why. To realize that we have overlooked such a fundamental aspect of social contagion not only opens up an entirely new subfield for future research, but forces us to reexamine much of what we already know.

Keywords:
Social Contagion, Social Learning, Social Networks, Polarization, Worldviews, Interdependent Diffusion, Experiment, Agent-Based Simulation

Main Body:

On June 17th, 2015, the U.S. Treasury announced that a portrait of a woman would appear on the ten-dollar bill *[24]*. The same day, news emerged of a mass shooting at a historically black church in Charleston, South Carolina *[25, 26]*. Within twenty-four hours both stories had spread throughout U.S. social media. It is reasonable to study the spread of these diffusants independently of one another, because the probability that an individual will share news of the shooting is not likely to be causally influenced by whether they have understood news about the $10 bill, and vice versa.

The day after the Charleston shooting, two other news items emerged. The first was a report that the shooter had been motivated by racial hatred symbolized in the Confederate flag. The second was a call to remove the Confederate flag from the South Carolina state capitol grounds *[25, 26]*. Even though these are distinct ideas, each spreading through social contagion, we cannot ignore one in trying to understand the diffusion of the other. If an individual has previously adopted the belief that the flag should be removed from the capitol grounds, they will be more likely to believe that the shooter's identification with the flag is politically relevant, and vice versa. Rather than being independent diffusants, these beliefs are *inter*dependent.

With good reason, nearly all social contagion research assumes that diffusants spread independently of one another. The independence assumption makes for parsimonious theory *[1-7,11-15,19-21]*, and it reduces the complexity and expense of experiments *[8-10, 16-18, 22, 23]*. The most influential authors on social contagion have used this assumption to study the effect of social network structure *[1-10]*, social reinforcement *[11-18]*, homophily *[19-21]*, and network rewiring *[21-23]* on contagion outcomes. Unfortunately, scholars assume independence between diffusants so frequently and to such productive ends that we generally forget we are doing so, and fail to question whether the assumption is appropriate.

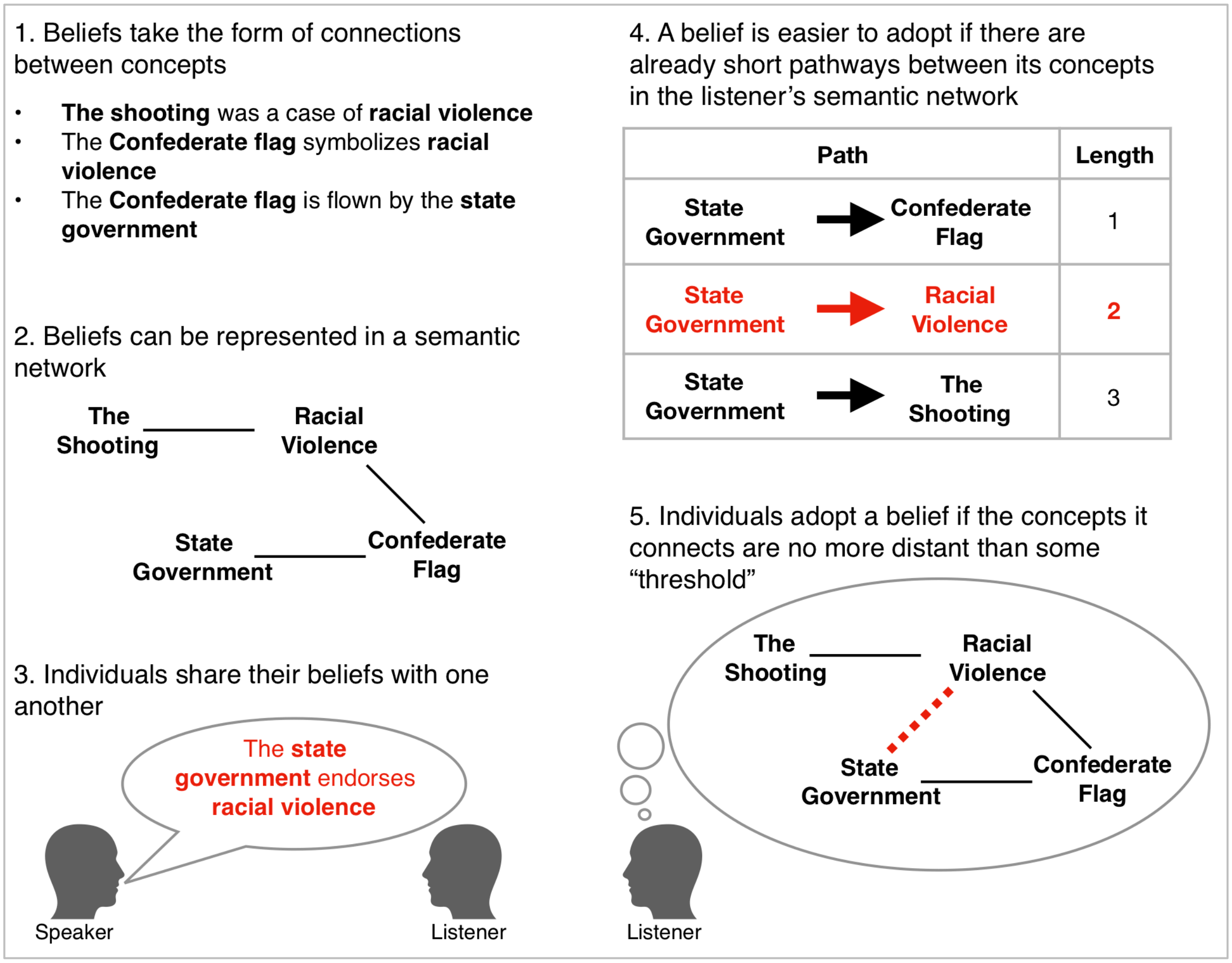
In contrast with independent diffusion, "interdependent diffusion" describes any social contagion process in which individuals' likelihood of adopting diffusant A is a function of their current state of adoption of B (C, D, …) *and* in which their likelihood of adopting B (C, D, …) is a function of their state of adoption of A. Very few theoretical models include any form of interaction between diffusants *[27-32]*, and none are empirically verified. It remains to be seen whether interdependence between diffusants 1) generates new sociological processes, 2) creates new observable outcomes, and 3) has practical consequences for communication and social policy. In short, does interdependence matter for the theoretical and empirical study of social contagion, or are our models of independent diffusion sufficient?

Interdependent diffusion can be studied through two theoretical lenses. First, we will observe the spread of a single belief as it participates in a process of "reciprocal facilitation" with other diffusants. Secondly, we will observe a pair of individuals as they exchange multiple beliefs with one another in a process we might call an "agreement cascade".

A focal belief spreads when an exposed individual's existing beliefs lead her to think it is true – that is, they "facilitate" her adoption of the focal belief. From its new position in the social network, the focal belief may then spread further, and subsequently facilitate the adoption of the beliefs that had previously supported its diffusion. This cycle creates a reinforcing feedback, such that when diffusants alternately create susceptibility to one another they can be adopted by more individuals than any single belief could have reached on its own. We see this “reciprocal facilitation” dynamic at work in beliefs about the confederate flag. The existing political conflict facilitated the spread of information about the shooter’s identification with the flag, and as it spread, news of the shooter’s identification brought more attention to the political conflict over the flag’s display at the state capitol.

We can use a simple simulation model to explore the macro-scale effects of the reciprocal facilitation process. In this model, individuals’ beliefs are represented using the “semantic network” abstraction borrowed from the cognitive science literature *[33-36].* As shown in Fig. 1, nodes in a semantic network represent concepts such as people, places, or activities, and edges represent the belief that two concepts are connected in some way (*e.g.* the *person* visited the *place)*. When an individual is exposed to new beliefs, she adopts those that connect two concepts that are already close together in her semantic network *[36, 37]*, as doing so does not require dramatic changes to her belief structure. The simplest representation of this tendency is that a simulated individual will adopt any belief that her neighbors possess, as long as the distance it spans in her existing semantic network is below some threshold[1].

[1] Other models assume relationships such as *“*belief *i* is compatible with *j*, but not *k”*. Avoiding the need to pre-specify compatibility gives confidence that any systematic variation in adoption can be attributed to the diffusion process rather than to the assumed relationships. The importance of this is outlined in the supplement.


1. Beliefs take the form of connections between concepts
The shooting was a case of racial violence
The Confederate flag symbolizes racial violence
The Confederate flag is flown by the state government
2. Beliefs can be represented in a semantic network
The Shooting
Racial Violence
State Government
Confederate Flag
3. Individuals share their beliefs with one another
The state government endorses racial violence
Speaker
Listener
4. A belief is easier to adopt if there are already short pathways between its concepts in the listener's semantic network
5. Individuals adopt a belief if the concepts it connects are no more distant than some "threshold"
The Shooting
Racial Violence
State Government
Confederate Flag
Listener

Path	Length
State Government → Confederate Flag	1
State Government → Racial Violence	2
State Government → The Shooting	3

Fig. 1. A minimal model of interdependent diffusion

Figure 2 shows a simulation of 60 agents in a random social network, who are assigned random beliefs, and who adopt beliefs that span up to two steps distance in their semantic networks[2]. First, we see that reciprocal facilitation allows the average number of people susceptible to a belief to grow endogenously with the average number of adopters (Fig. 2A). As a comparison, black curves show how the same beliefs would diffuse if they did not interact. In the independent case, beliefs can only be

[2] The results of the simulation are qualitatively similar with other social networks or decision rules. See the methods section for more details.

adopted by individuals who are susceptible at the start, and so they spread less widely in the population. Put another way: to explain the same level of final adoption, models of independent diffusion need to assume significantly more initial susceptibility to each belief.

It might be reasonable for us to ignore interdependence (and instead assume more widespread initial susceptibility) were it not for the second effect of reciprocal facilitation. When diffusants are independent, the population initially susceptible to a belief is an excellent predictor of who will eventually adopt it (Fig. 2B). However, when beliefs interact, individuals can acquire susceptibility through adoption of other beliefs, such that initial susceptibility is no longer a strong predictor of eventual adoption. For example, imagine that a "focus group" is selected from our artificial population before the simulation starts, and that amongst this group belief A is adopted by 25% more people than belief B. Unsurprisingly, if we simulate the spread of these beliefs independently of one another, belief A will be more popular than B in over 99% of cases. On the other hand, when we simulate interdependent diffusion, A is adopted by more people than B only 57% of the time – just slightly better than chance.

The popularity of interdependent beliefs is hard to predict because the reciprocal facilitation process does not support the diffusion of all beliefs to the same degree. Instead, it preferentially amplifies those that are connected to other widely adopted beliefs. Figure 2C shows that a belief's popularity becomes correlated with that of the most popular belief adjacent to it. This is because a belief that is supported by many popular beliefs will have many opportunities to diffuse, and then to facilitate the adoption of other closely related beliefs. Conversely, a belief that is connected only to unpopular supporting beliefs will have trouble reaching even the few individuals who are susceptible to adopting it.

As a result, patterns emerge when individual semantic networks are aggregated to the level of the population. After interdependent diffusion, clusters of mutually-supporting beliefs are held by large fractions of the population, and they dictate which new beliefs can be adopted (Fig 2D). When we observe these belief structures in the real world *[25]* we call them “worldviews”, and we assume that they reflect an underlying truth or natural grouping of especially compatible beliefs. To the contrary, this model of interdependent diffusion suggests that worldviews are accidents of history: different worldviews emerge each time we run the simulation. Models of independent diffusion, on the other hand, must explain macro-level belief structures by reference to some external cause.

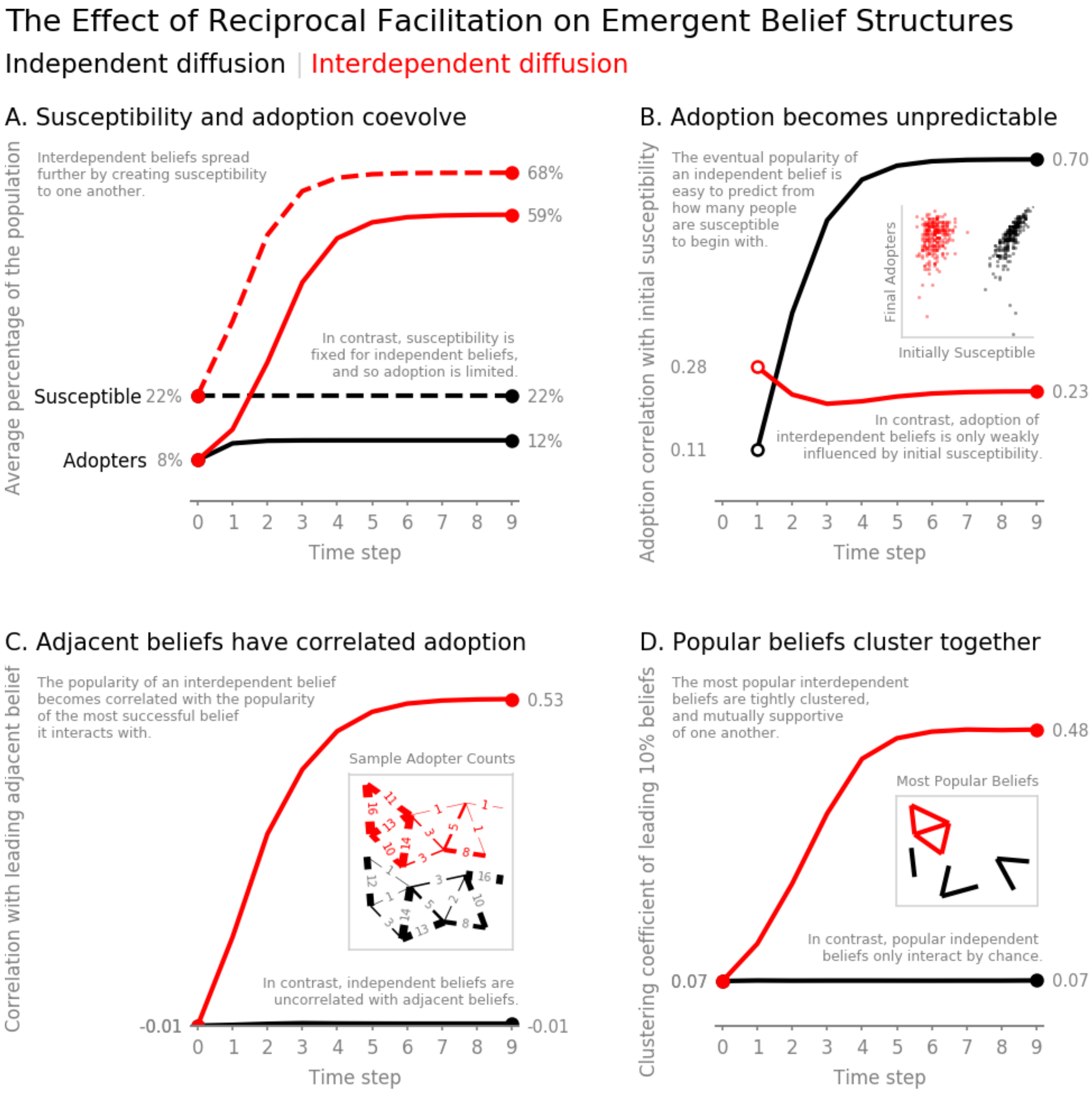


Fig. 2: The effect of reciprocal facilitation on emergent belief structures.

Switching lenses, we now focus on a pair of neighboring individuals, and observe a process of “agreement cascades”. When two people exchange beliefs, they become more similar to one another. Because their existing beliefs influence the way people respond to new diffusants, shared beliefs make the two individuals more likely to adopt (or reject) the same new beliefs in the future. As a result, they become more similar still, regardless of any preference to align or distinguish themselves from one another.

In our simulation, agreement cascades lead to the emergence of increasingly self-similar ideological “camps” that expand members’ access to beliefs they can hold in common, and filter out beliefs that would set them apart from one another (Fig. 3A). Differences between camps are amplified as existing beliefs drive dissimilar individuals to adopt beliefs from different parts of the semantic space.

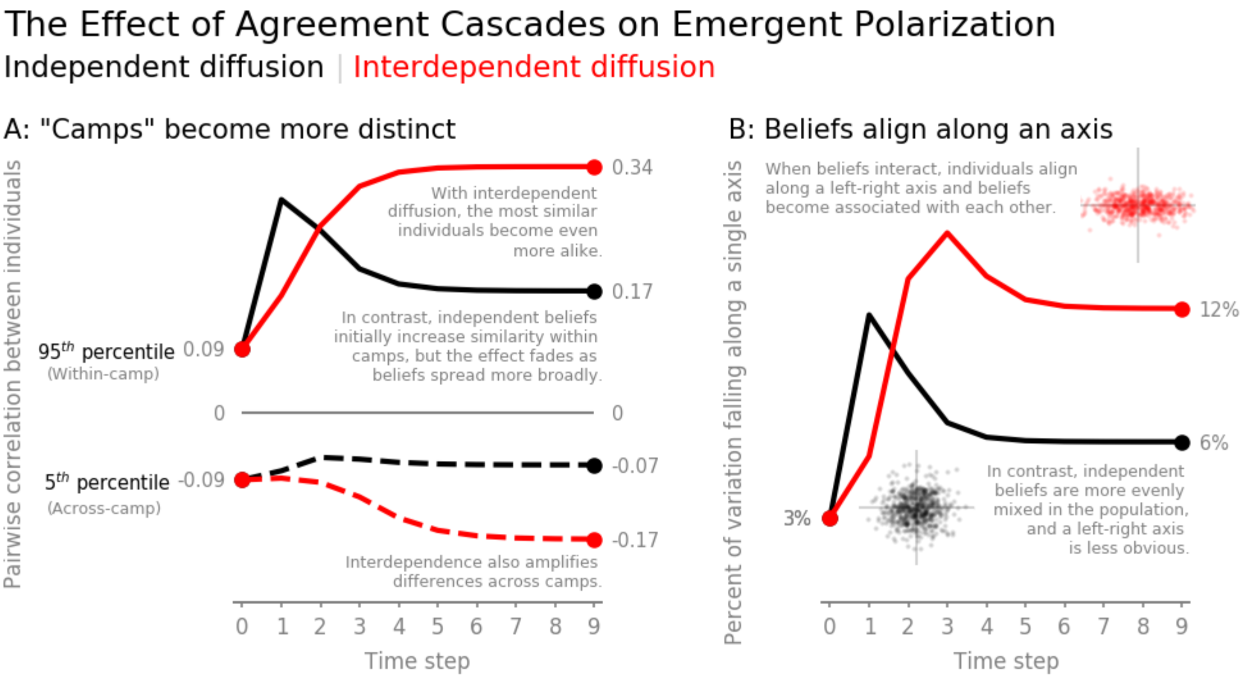


Figure 3: The effect of agreement cascades on emergent polarization

As individuals organize into camps, it becomes easier to predict each person’s position on one belief from their position on other beliefs. For example, belief A may co-occur with belief B 70% of the time, but co-occur with belief C only 10% of the time. This compresses the population’s variation in the space of beliefs into a few principal axes (e.g. liberal-conservative, or libertarian-populist). As a result, differences between individuals can be increasingly described by their position along a “left-right axis”, as seen in Fig. 3B, with individuals at any point on the axis relatively similar to one another. When beliefs diffuse independently of one another, there is less alignment along the left-right axis and more variation between individuals at any given point.

This simulation predicts that interdependence between diffusants will create at least two new outcomes of social contagion, but (as a deliberately simplified model) it may mischaracterize human behavior in important ways. To test whether the predictions hold when actual people exchange realistic beliefs, I conducted a fully preregistered[3], randomized controlled experiment with 2400 participants. In an online laboratory context, I systematically varied the level of interaction between otherwise identical diffusants in identically constructed populations, and measured emergent polarization in individuals’ behavior and self-reported beliefs.

The experiment took the form of a “detective game” in which participants were given clues to solving a burglary, and were asked to share any promising leads with their neighbors in a 20-person social network. A subset of clues (~30%) was common to both independent and interdependent experimental conditions, and was used for analysis. The remaining clues either created connections between the analysis clues (interdependent condition) or gave additional, non-interacting details (independent condition). Unknown to participants, clues could equally implicate any suspect or

[3] The preregistration can be found at https://osf.io/239ns

burglary method and were seeded equally in the population. See the methods section for details.

Despite the symmetry of the setup, and the fact that there was *by design* no solution to the mystery, participants came to strongly-held beliefs about which suspect was guilty and how they performed the crime. On average, participants felt that the most likely suspect was almost 50% more likely to be involved in the crime than the least likely suspect, and over half of participants reported confidence in at least one part of their solution of 95% or greater. Qualitatively, the “solutions” that participants surmise are clusters of mutually-reinforcing beliefs that arise by chance, and persist though their influence on which new beliefs are adopted – much as we described the emergence of worldviews in the simulations above.

In a testament to the real-world relevance of belief interaction, the experiment was not able to perfectly operationalize the assumption of independence between diffusants, despite ideal laboratory conditions. For example, even though the independent clue set contained no explicit links between suspects, participants could draw implicit connections between the guilt of one suspect and the presumed innocence of another. This imperfect control means that measured differences between independent and interdependent conditions are likely to underestimate the true effect of belief interaction.

Nevertheless, the results of this experiment support the above theoretical predictions. Figure 4A shows that interdependence measurably increased the population’s alignment along a left-right axis among both behavioral (+2.2% p=.013) and self-reported (+2.8% p=.022) measures of belief. While not all measures were significant, camps were more self-similar (behavioral measures +.025 p=.014, Fig. 4B) and more distinct from one another (self-report measures -.035 p=.058, Fig. 4C) in the interdependent condition than in the independent condition.

To gauge whether the effect of interdependence is large enough to be worth attention when compared to other drivers of polarization, I ran a parallel experimental condition that varied the social network structure between non-polarizing and polarizing extremes. Real-world social networks lie somewhere between these two extremes, and so this manipulation creates an *a fortiori* comparison effect. If the effect of interdependence is nontrivial in this comparison, then we have confidence that it is meaningful in the real world. In the statistically significant measures of this experiment, interdependent diffusion creates an effect between 16% and 39% of the *a fortiori* reference, as shown in Fig. 4. We should always use caution when generalizing effect sizes from laboratory groups to large-scale social networks. However, this comparison suggests that interdependence plays an unignorable role in social contagion, and that scholars' almost exclusive focus on network structure over belief interaction is out of proportion to the relative importance of the two effects.

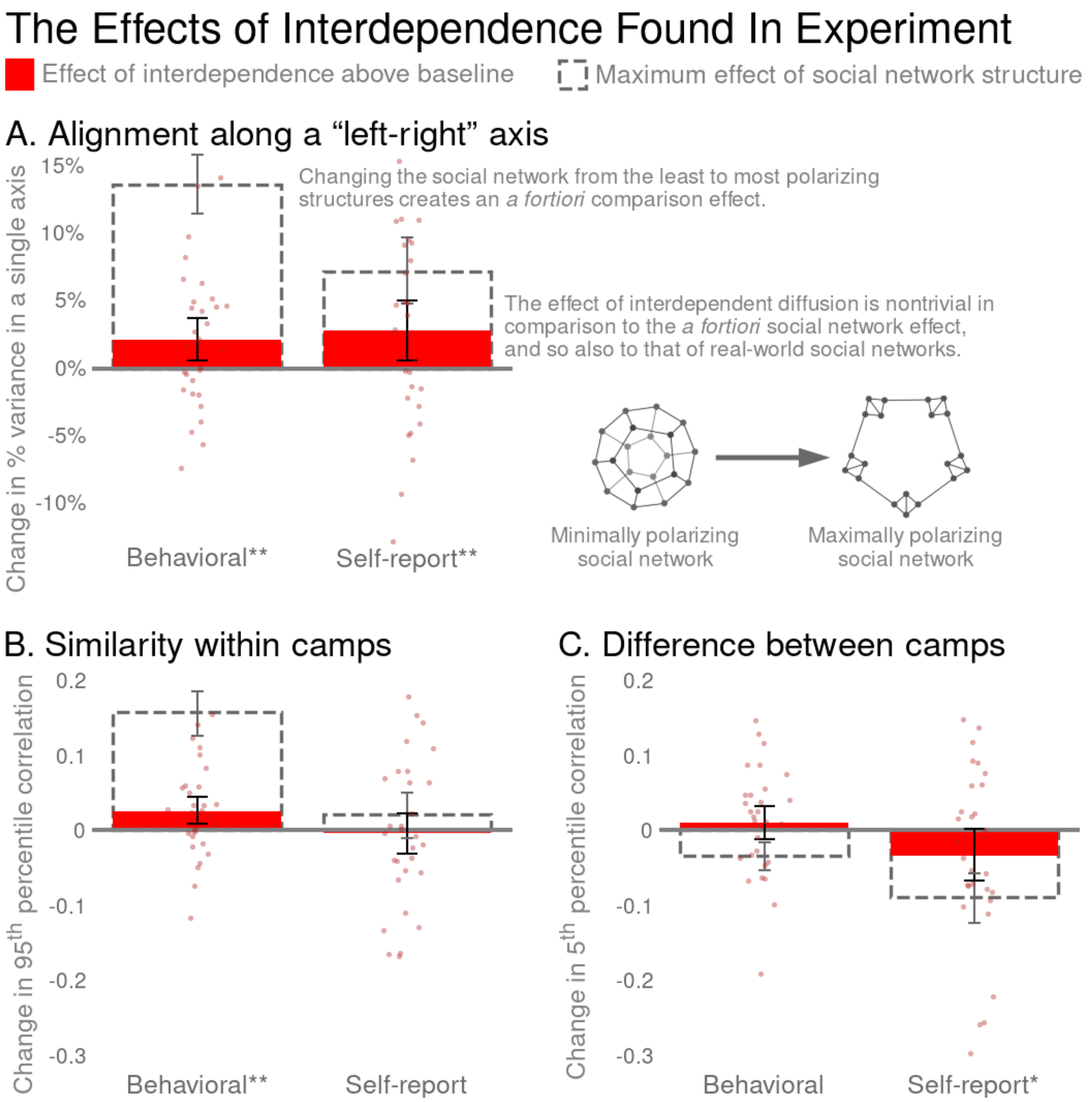


*Figure 4. The effects of interdependent diffusion observed in experiment. *p<.1 **p<.05. Error bars show 90% CI. n=30 pairwise comparisons between social networks.*

While it cannot explore all of the implications of interdependent diffusion, this paper demonstrates that belief interaction enables at least two new sociological processes that influence polarization and the emergence of worldviews. The social contagion literature is replete with studies that explain polarization as a consequence of social network structure *[38-39]*, homophily *[40-41]*, complex contagion *[42-43]*, etc. This paper shows that when beliefs interdepend, none of these explanations are necessary; polarization is a natural consequence of diffusion itself. Rather, what requires explanation is consensus. Likewise, we have ample research asking how fringe beliefs and conspiracy theories emerge and persist *[42-44]*. Interdependent diffusion suggests that they are a natural consequence of social contagion. What should surprise us is that populations are occasionally able to find the truth.

This simulation and experiment remind us that the simplifying assumption of independence between diffusants is just that – an assumption. Despite its ubiquity, the assumption should be carefully made and frequently challenged. With luck we will find that most of what we know about diffusion is robust to interaction between diffusants. We may also find that relaxing the assumption of independence helps us explain new sociological phenomenon and better understand social contagion.

1. M. Granovetter. The Strength of Weak Ties. *Am. J Sociol.* **78**, 1360-1380 (1973).
2. R. Burt. Structural Holes: the social structure of competition. *Harvard University Press* (1992).
3. D. Watts and S. Strogatz. Collective Dynamics and Small World Networks. *Nature.* **393,** 440-442 (1998).
4. D. J. Watts, A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci. U.S.A.* **99** 5766-5771 (2002).
5. R.Reagans and B McEvily. Network structure and knowledge transfer: The effects of cohesion and range. *Admin Sci Quarterly.* **48,** 240-267. (2003).
6. D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the Spread of Influence through a Social Network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2003.

7. D. Lazer and A. Friedman. The network structure of exploration and exploitation. *Admin Sci Quarterly.* **52** 667-694 (2007).
8. J. Travers and S. Milgram. An Experimental Study of the Small World Problem. *Sociometry.* **32** 425-443 (1969).
9. M. Kearns, S. Suri, N. Montfort. An experimental study of the coloring problem on human subject networks. *Science*. **313**. 824-827 (2006).
10. S. Suri and D. Watts. Cooperation and Contagion in Web-Based, Networked Public Goods Experiments. *PLoS ONE* **6** e16836 (2011)
11. G. Le Bon. The Crowd: A study of the popular mind. T Fisher Unwin. (1897).
12. M. DeGroot. Reaching a consensus. *J Am Stat Assoc. **69**. 118-121. (1974)*
13. T. Schelling, "Sorting and Mixing" in *Micromotives and Macrobehavior*. (Norton, Toronto, 1978) chap. 4.
14. M. Granovetter, Threshold Models of Collective Behavior. *Am. J Sociol*. **83**, 1420-1443 (1978).
15. D. Centola, M. Macy, Complex Contagions and the Weakness of Long Ties. *Am. J Sociol.* **113** 702-734 (2007).
16. M. Salganik, M. Dodds, and D. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. Science. **311** 854-856 (2006).
17. J. Lorenz et al. How social influence can undermine the wisdom of crowd effect. *PNAS* **108** 9020-9025 (2011).
18. L. Muchnik, S. Aral, and S. Taylor. Social Influence Bias: A Randomized Experiment. *Science* **341** 647-651 (2013).
19. C. Shalizi and A. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociol. Methods Res.* **40** 211-239. (2011).
20. B. Golub and M. Jackson. How homophily affects the speed of learning and best-response dynamics. *Q J Econ.* **127** 1287-1338 (2012).
21. N. Christakis and J. Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Stat. in Medicine.* **32** 556-577. (2013).
22. D. Rand, S. Arbesman, and N. Christakis. Dynamic Social Networks Promote Cooperation in Experiments with Humans. *PNAS* **108** 19193-19198 (2011).
23. A. Almaatouq et al. Adaptive social networks promote the wisdom of crowds. *PNAS.* **117** 11379-11386. (2020)
24. J. Calmes. "Treasury Says a Woman's Portrait Will Join Hamilton's on the $10 Bill. *New York Times,* 18 June 2015, p. A20.

25. J. P. Houghton *et al.* Beyond Keywords: Tracking the Evolution of Conversational Clusters in Social Media. *Sociol. Methods Res.* **48**, 588-607 (2019).
26. J. Hawes. Grace will lead us home: The Charleston Church Massacre and the Hard, Inspiring Journey to Forgiveness. *St. Martin's Press*, *New York*. (2019).
27. D. Baldassarri, P. Bearman, Dynamics of political polarization. *Am. Sociol. Rev.* **72** 784-811 (2007).
28. D. DellaPosta *et al.*, Why do liberals drink lattes? *AM. J. Sociol.* **120** 1473-1511 (2015).
29. N. E. Friedkin *et al.*, Network science on belief system dynamics under logic constraints. *Science.* **345**, 321-326 (2016).
30. A. Goldberg, S. Stein, Beyond Social Contagion: Associative Diffusion and the Emergence of Cultural Variation. *Am. Sociol. Rev.* **83** 897-932 (2018).
31. S. E. Parsegov *et al.*, Novel Multidimensional Models of Opinion Dynamics in Social Networks. IEEE Trans. Automat. Contr. **62** 2270-2285 (2017).
32. F. Xiong *et al.* Analysis and application of opinion model with multiple topic interactions. *Chaos.* **27** 083113 (2017).
33. A. M. Collins, E. F. Loftus, A Spreading-Activation Theory of Semantic Processing. *Psychol. Rev.* **82** 407-438 (1975).
34. M. Steyvers, J. B. Tenenbaum, The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cogn. Sci.* **29** 41-78 (2005).
35. R. J. Brachman, "On the epistemological status of semantic networks" in *Associative Networks.* 3-50 (1979).
36. M. Schilling, A 'small-world' network model of cognitive insight. *Creat. Res. J.* **17** 131-154 (2005).
37. R. S. Nickerson. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* **2** 175–220 (1998).
38. A. Flache, and M. W. Macy. Small worlds and cultural polarization. *J of Math Sociol* **35** 146-176 (2011).
39. M. Del Vicario, G. Vivaldo, A. Bessi, *et al.* Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Sci Rep* **6,** 37825 (2016).
40. P. Dandekar, A. Goel, and D. Lee. Biased assimilation, homophily, and the dynamics of polarization. *PNAS* **110**, 5791-5796. (2013).
41. V. Vasconcelos, S. Levin, and F. Pinheiro. Consensus and polarization in competing complex contagion processes. *J Royal Society Interface*, *16*(155), p.20190196. (2019)

42. D. Spohr, Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Bus Inform Rev* **34**, 150-160 (2017).

43. P. Törnberg, Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, **13**, e0203958. (2018).

44. G. Pennycook, and D. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188,** 39-50. (2019).

Methods:

Simulation:

In the simulation presented in Figs. 2 and 3, the social network is a connected Erdős–Rényi (G_{nm}) random graph with 60 agents, each with an average of 3 neighbors. Each agent is initialized with 25 beliefs (edges) selected randomly from the 300 edges available in a complete semantic network with 25 concepts (nodes). These values ensure good coverage of beliefs in the population, while individual semantic networks are initially sparse. Random seeding ensures that the simulation starts without polarization or systematic variation in belief popularity, and also that the social network structure itself does not contribute to polarization. Because beliefs are drawn from a complete semantic network, there is no natural belief structure around which polarization may nucleate. Results are qualitatively similar with other types and sizes of social networks, and different sizes and seeding densities of initial semantic networks, so long as there are enough beliefs seeded initially for diffusion to occur, and not so many that adoption is complete.

In each step, individuals are selected in random order, and update their beliefs by incorporating into their semantic networks all beliefs (edges) that their neighbors possess and they are susceptible to adopting. In the interdependent case, individuals are susceptible to any belief with an existing path length of 2 (i.e. that closes a triangle) in their semantic networks at the current time. In the comparison (independent) case for Fig 2A, a random selection of the population is defined to be susceptible to each belief in the same proportion as are initially susceptible to the belief in the interdependent diffusion condition. In Figs 2B-D and Fig 3, a random selection of susceptible individuals is made in proportion to the *final* number of susceptible individuals in the interdependent diffusion case. As a result, for all graphs other than 2A, a histogram of the extent of diffusion of each belief is approximately the same under both independent and interdependent treatments. This ensures that the subsequent presentation of results reflects purely the effect of interdependence between diffusants, and not the effect of different levels of adoption in the compared populations. Results are qualitatively similar when calculated based upon the initial susceptibility. All code necessary to run the simulation is available in the supplement and at https://github.com/JamesPHoughton/interdependent-diffusion.

Measures

The measures presented in Figs. 2A-C are averaged over all beliefs in the simulation, and over 20,000 simulations. The measures in Figs. 2D and 3 are population-level measures averaged over 20,000 simulations. This volume of simulations is an order of magnitude beyond the point at which noise affects the result.

The measure of the susceptible population in Fig. 2A (and all discussion of the susceptible population) represents all individuals who would adopt the belief if exposed to it, along with all of the individuals who have already adopted the belief. Fig. 2B measures the correlation between new adoption and those who are initially susceptible to the belief but do not start with it. As this has no meaningful value at t0, the curve is drawn from t1-t9.

Figure 2D uses the clustering coefficient of a semantic network constructed from the most popular 10% of beliefs as a demonstration that the most popular beliefs are mutually interrelated, and not merely all related to a single leading belief (e.g. a star or barbell pattern). Clustering only makes sense when beliefs are conceptualized as a semantic network. Other conceptualizations of belief interaction might prefer to plot the number of top decile beliefs that each top decile belief interacts with. This measure gives essentially the same result (i.e. large fractional growth over time in the interdependent case, with no change from randomness in the independent case) but fails to capture the mutual interrelatedness indicated by the clustering coefficient. The measure is generally insensitive to the specific threshold used to define a 'popular' belief for any thresholds between about 5% and 40%. See the supplement for sensitivity analysis.

There are many complex measures of polarization in the literature *[27,28,30, 38-44 for a sample],* which generally attempt to represent three basic intuitions. First, that individuals within the same ideological camp come to be more similar to one another. Secondly, that individuals in different ideological camps become more dissimilar to one another. Lastly, that an individual's position on one dimension of belief becomes informative of their position on other dimensions. As my purpose is not to identify camps and their members, but to suggest that one set of conditions is more generative of polarization than another, these measures add more complexity than value. Instead I report heuristic measures characterizing the above three intuitions.

A simple and reproducible way to assess the similarity of individuals within an ideological camp – absent exogenous labels such as demographic or party – is to measure the similarity between all pairs of individuals and define a certain percentile as belonging to the same ideological camp. The more exclusive we are (i.e. the higher the percentile), the more conservative the claim that these represent "within-camp" relationships. To define across-camp similarity, we can choose a percentile that (conservatively) represents relationships between individuals in different ideological camps. In Figs. 3A, 4B, and 4C, I use the 95th and 5th percentiles. See the supplement for sensitivity analysis.

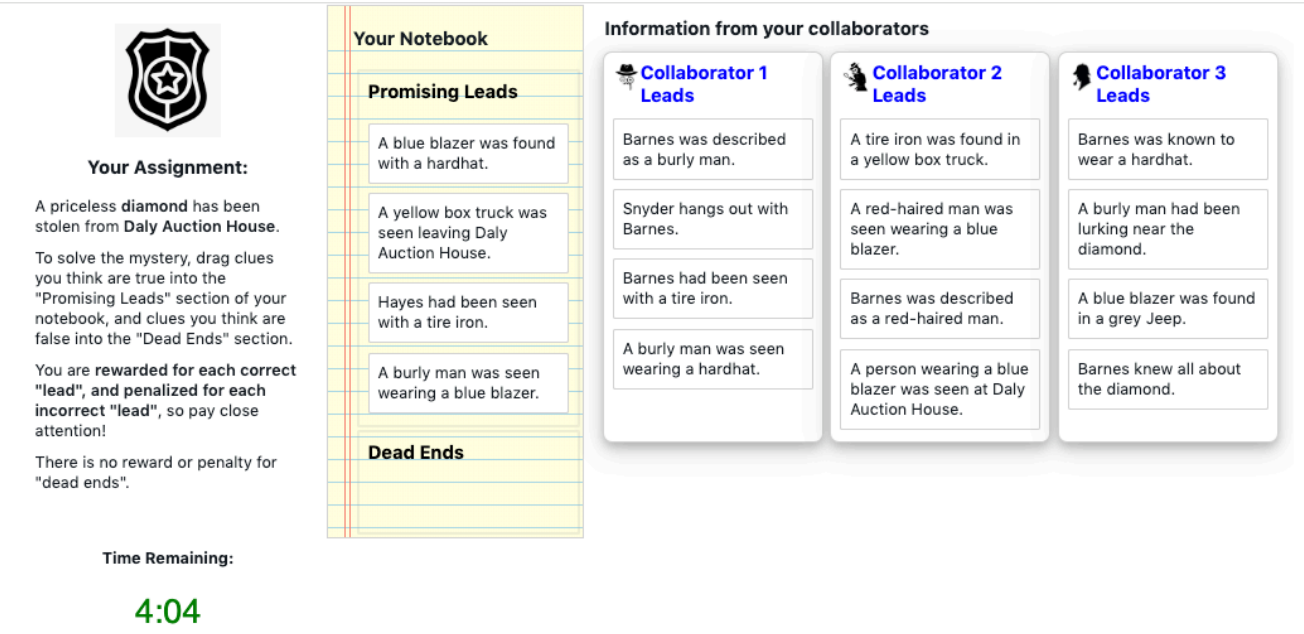
Figs. 3B and 4A measure belief alignment as the percent of variation between individuals that can be explained by the best fitting axis in the space of possible beliefs, using singular-value decomposition. In Fig 3B, the original feature space has one dimension for each belief in the simulation (300), and points representing each individual's position in that feature space (60) according to the beliefs they have adopted. The inset graphs are exaggerated and show a larger population to illustrate how a component can explain more or less variation. In Fig 4A, the feature space has

22 dimensions for the behavioral measures, and 11 dimensions for self-report measures (described below), and points representing 20 individuals.

Experiment

Over the course of eight days in July 2020, I recruited 2768 U.S. and Canadian Mechanical Turk workers under the criteria that they were 18+ years old, and have completed 100+ HITs with a 90%+ approval rating. Of these, 2400 completed training and were randomized into 20-person social networks. Each network was assigned to one of four (matched) experimental conditions. Each condition had n=30 samples, described below. This sample size was set by budget constraints. The participant population was 45% female; mean 37.1 years old; 27% high-school, 49% bachelors, 16% masters+ graduates. 96.8% of players who completed training went on to complete all steps of the experiment, with less than 0.4% difference in dropout between conditions. Participants were paid $0.10 for showing up, $1 for training in how to use the interface, $1 to participate, up to $1 individual performance bonus, and up to $1 for the performance of their team. The average payout was just over $4, and the experiment took about 20 minutes. The preregistration for this experiment included all code necessary to generate test conditions, implement the game using Empirica *[45]*, conduct the experiment, and process the resulting data. This code, along with anonymized timestamped data for every player action and processed experiment data, is available for further analysis at https://github.com/JamesPHoughton/detective-game-interdependent-diffusion.

In this experiment, participants were asked to find a solution to a mystery by identifying a burglar's name, description, clothing, burglary tool, and getaway vehicle. Participants were seeded with four clues to the mystery, and were incentivized to sort those clues into "Promising Leads" and "Dead Ends". When a participant categorized a clue as a promising lead, it was immediately shared with their three neighbors in the 20-person social network. The primary experiment interface is shown in Fig. 5, and the participants' experience is fully detailed in the supplement.



Your Assignment:

A priceless **diamond** has been stolen from **Daly Auction House**.

To solve the mystery, drag clues you think are true into the "Promising Leads" section of your notebook, and clues you think are false into the "Dead Ends" section.

You are **rewarded for each correct "lead", and penalized for each incorrect "lead"**, so pay close attention!

There is no reward or penalty for "dead ends".

Time Remaining:

4:04

Your Notebook

Promising Leads

A blue blazer was found with a hardhat.

A yellow box truck was seen leaving Daly Auction House.

Hayes had been seen with a tire iron.

A burly man was seen wearing a blue blazer.

Dead Ends

Information from your collaborators

Collaborator 1 Leads

Barnes was described as a burly man.

Snyder hangs out with Barnes.

Barnes had been seen with a tire iron.

A burly man was seen wearing a hardhat.

Collaborator 2 Leads

A tire iron was found in a yellow box truck.

A red-haired man was seen wearing a blue blazer.

Barnes was described as a red-haired man.

A person wearing a blue blazer was seen at Daly Auction House.

Collaborator 3 Leads

Barnes was known to wear a hardhat.

A burly man had been lurking near the diamond.

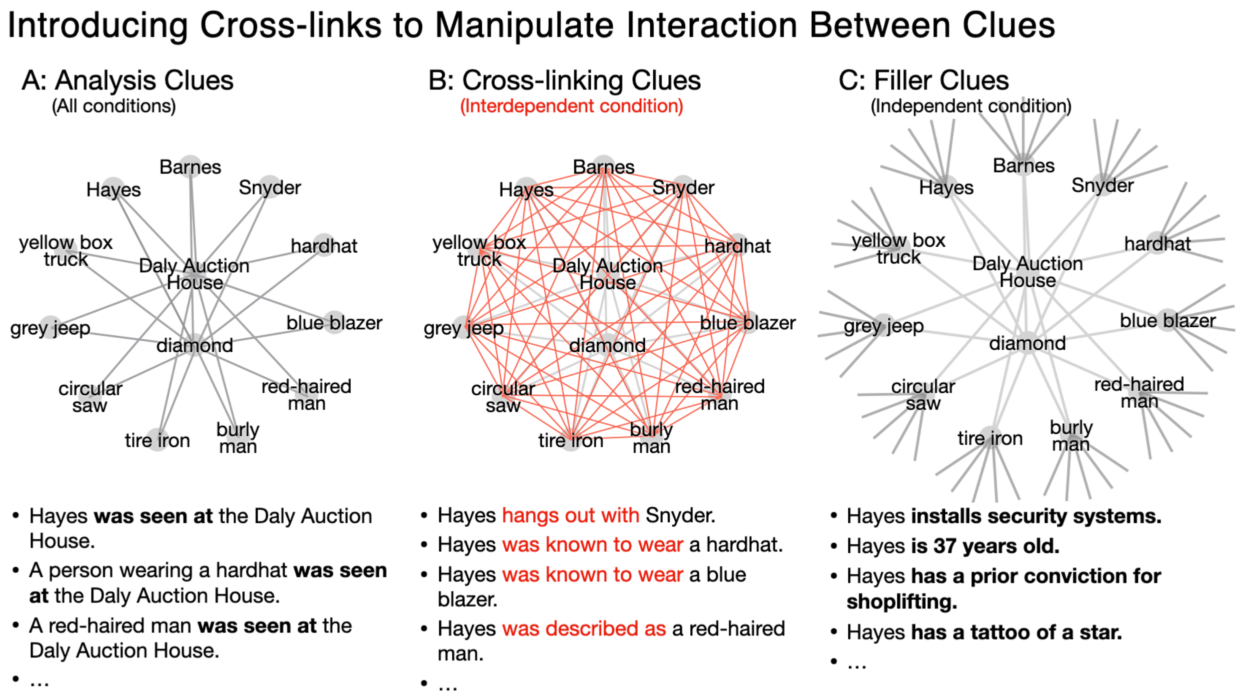
A blue blazer was found in a grey Jeep.

Barnes knew all about the diamond.

Fig. 5. The primary user interface of the "Detective Game". Clue cards can be dragged into categories in the player's "Notebook". Promising Leads are immediately shared with three neighbors, who can drag them into their own notebooks.

In the theory-building simulations in this paper, agents can be programmed to pay attention to interactions between beliefs in an "interdependent" world, and ignore those interactions in an identical “independent” world. Human beings, on the other hand, are wired to see connections and relationships between ideas, and so it is impossible to conduct a perfect test in which *all* clues interact in one experimental condition, while *none* of those clues interact in another. Instead, I partitioned the clues such that some were common to both independent and interdependent conditions and used for analysis (i.e. “analysis clues”), while others varied across conditions to manipulate the level of interaction between analysis clues. In each game, 22 analysis clues linked the crime scene and stolen object to each of 3 suspects, 2 physical descriptions, 2 items of clothing, 2 tools, and 2 vehicles, as illustrated in Fig. 6A. In the interdependent condition, 55 additional “cross-linking clues” connected all of these elements of the mystery (i.e. suspects, vehicles, etc.) to one another (Fig. 6B), to create a complete semantic network of clues. In the independent condition, 55 “filler clues” took the place of cross-linking clues to break the relationships between elements (Fig. 6C). These filler clues ensured that the two conditions had the same total number of clues, and that the structure of clues was as similar as possible between conditions.

Clues were extensively pre-tested to minimize bias from the outside world, such that each element was perceived to be equally likely to be involved in a burglary absent other information. Sets of clues were then randomly generated from the pool of pretested elements, representing over 225 billion possible mysteries. At the start of the game, each clue was present in exactly one player’s notebook. As a result, beliefs spread on a level playing field, such that *a priori* none should be expected to diffuse more than any other. Further information can be found in the “Clue Generation” section of the supplement.



- Hayes **was seen at** the Daly Auction House.
- A person wearing a hardhat **was seen at** the Daly Auction House.
- A red-haired man **was seen at** the Daly Auction House.
- …

- Hayes hangs out with Snyder.
- Hayes was known to wear a hardhat.
- Hayes was known to wear a blue blazer.
- Hayes was described as a red-haired man.
- …

- Hayes **installs security systems.**
- Hayes **is 37 years old.**
- Hayes **has a prior conviction for shoplifting.**
- Hayes **has a tattoo of a star.**
- …

Fig. 6: Clues are designed such that in the interdependent condition, “analysis” clues are connected to one another by cross-linking clues; while in the independent condition, analysis clues remain disconnected from one another.

In addition to the two types of clue structures, the experiment also included two types of social networks, in both of which 20 players were connected to exactly three neighbors each (Fig. 4 inset). The first network was a “dodecahedral” network, in which none of a player’s neighbors were directly connected to any other (0% clustering), and the average network distance between individuals was short (2.6 steps[4]). We should expect to find very little polarization in this network, as information can diffuse across

[4] The maximum average distance between 20 individuals in a connected network is 7

the network readily, and coordination among subgroups is impeded by the lack of mutual connections. The second social network was a “regular connected caveman” structure, in which neighbors shared 50% of their remaining contacts in common (50% clustering), and there are large average distances between individuals (4 steps). We should expect to find high levels of polarization in this network regardless of the level of interaction between clues, as strong clustering makes it easy for subgroups to converge on a shared set of clues, and long average path lengths make it harder for information to spread between camps. Further information can be found in the “Social Network Structure” section of the supplement.

Together these manipulations create four separate conditions. The dodecahedral network and independent clue set formed a baseline condition. Contrasted with the baseline, experimental conditions measured the effect of interdependence, the effect of social network structure, and the effect of both manipulations combined. Blocks of four simultaneous games were constructed with one game in each condition. To guard against latent external biases, each block used a different randomly-generated set of clues. Clue assignments varied as little as possible within each block, such that the clues assigned to a particular network position in one game corresponded to those assigned to the same network position in the other three games. Upon completing training, participants were randomly assigned to positions within a block, blind to their treatment condition. I treated games within each block as matched samples. As there is a strong theoretical prior for the direction of each effect, I used one-sided pairwise t-tests to assess the differences between each experimental condition and the baseline.

Bootstrapped confidence intervals in Fig. 4 include 90% of resampled cases, to correspond with the one-sided pairwise t-tests at p<.05 significance level.

The measures of polarization described in the above simulation were operationalized to reflect participants' behavior, and also their self-reported opinions. "Behavioral" measures were constructed from each participant's final categorization of the 22 analysis clues (i.e. the clues that were common to both the independent and interdependent condition), reflecting the cumulative choices that the participants had made throughout the game. Following the game, participants estimated the likelihood that each suspect, vehicle, etc. was involved in the crime. "Self-report" measures were constructed from these 11-item assessments, reflecting how participants internalized the information they encountered to create opinions. Finally, participants were asked to rate their confidence in their overall solution, and the level of consensus they perceived in their team.

The "self-reported" beliefs of experiment participants fall on a continuous scale from 0 to 100, and so it is natural to use Pearson's correlation on the vectors of individuals' beliefs to assess similarity between participants. This measure has the advantage of being easily interpretable and having a well-defined range that is independent of the number of features in the vector of attributes being compared, and the negative region of which can be interpreted as expressing dissimilarity. To assess the similarity of the binary "behavioral" data I use the Phi coefficient, an analogous measure to Pearson's correlation with the same interpretable range.

The behavioral measures in the experiment are sensitive not only to interdependence and network structure but also to the average level of diffusion of beliefs. To minimize noise due to differences in the level of activity between games, each of the behavioral measures is assessed compared to what would be expected due to chance, keeping the number of adopters of each clue and the number of clues adopted by each participant fixed. This correction was designed in simulation and preregistered.

In addition to the results presented in Fig. 4, I also computed an interaction effect of network structure and interdependence, and an effect of interdependence on players' confidence in their result and their estimate of consensus among their teammates. These results are presented in the supplement.

References:

45. P. DiMaggio, J. Evans, B. Bryson. Have Americans Social Attitudes Become More Polarized? *Am. J Sociol.* **102** 690-755 (1996).
46. P. Dandekar, A. Goel, and D. Lee. Biased assimilation, homophily, and the dynamics of polarization. *PNAS* **110** 5791-5796.
47. J. Becker, E. Porter, and D. Centola. The wisdom of partisan crowds. *Proc National Acad Sc*i. **116**, 10717–10722 (2019).
48. I. Permanyer. The conceptualization and measurement of social polarization. J Econ Inequal 10, 45–74 (2012).
49. D. Baldassarri and A. Gelman. Partisans Without Constraint: Political Polarization and Trends in American Public Opinion. *Ssrn Electron J* **114**, 408–446 (2008).
50. K. T. Poole H. Rosenthal. The Polarization of American Politics. *J Politics* **46,** 1061–1079 (1984).
51. J.M. Esteban and D. Ray. On the Measurement of Polarization. *Econometrica* **62**, 819 (1994).
52. A. Almaatouq, J. Becker, J.P. Houghton, N. Paton, D.J. Watts, and M.E. Whiting. Empirica: a virtual lab for high-throughput macro-level experiments. *arXiv preprint arXiv:2006.11398*. (2020).

Acknowledgments: Special thanks to Abdullah Almaatouq, Sinan Aral, Carolyn Fu, Hazhir Rahmandad, David Rand, Ray Reagans, Hagay Volvovsky, Duncan Watts, and the System Dynamics and Economic Sociology groups at MIT Sloan for their guidance.