# The Dynamics of Retraction in Epistemic Networks

Travis LaCroix[a,b], Anders Geil[a,c], Cailin O'Connor[a]

[a] *Department of Logic and Philosophy of Science*
*University of California, Irvine*

—

[b] *Mila*
*Québec Artificial Intelligence Institute*

—

[c] *Department of Computer Science*
*Columbia University*

**Abstract**

Sometimes retracted or thoroughly refuted scientific information is used and propagated long after it is understood to be misleading. Likewise, sometimes retracted news items spread and persist, even after it has been publicly established that they are false. In this paper, we use agent-based models of epistemic networks to explore the dynamics of retraction. In particular, we focus on why false beliefs might persist, even in the face of retraction. We find that in many cases those who have received false information simply fail to receive retractions due to social dynamics. Surprisingly, we find that in some cases delaying retraction may increase its impact. We also find that retractions are most successful when issued by the original source of misinformation, rather than a separate source.

*Email addresses:* tlacroix@uci.edu (Travis LaCroix[a,b]), a.geil@columbia.edu (Anders Geil[a,c]), cailino@uci.edu (Cailin O'Connor[a])

*Draft of April 16, 2020*

## 1. Introduction

Over the course of a decade, Scott S. Reuben, an influential American anaesthesiologist, published a series of articles examining the role of cyclooxygenase-2 specific inhibitors in controlling post-operative pain following orthopaedic surgery (Buckwalter et al., 2015). In 2010, though, Reuben was convicted of fraud for accepting grants from drug companies to perform clinical trials, but then fabricating the data and publishing 'results' without having conducted trials.[1] An investigation determined that at least 21 of Reuben's articles contained fabricated data; all of these articles were subsequently retracted (Shafer, 2015).

Before retraction, the articles had collectively obtained nearly 1200 citations. By 2014, however, a case study found that roughly *half* of these articles continued to be cited consistently, and only 1/4 of the citing articles clearly stated that Reuben's work had been retracted (Bornemann-Cimenti et al., 2016). It seems remarkable that scientific findings of this sort would be so widely used in the literature after being withdrawn as fraudulent, but this is by no means an isolated case. Studies of retracted results have found that they are often widely cited after retraction. Moreover, Cor and Sood (2018) find that 91% of these post-retraction citations are approving of the original research.[2]

In this paper, we examine how false information perpetuates, even in light of correction. To explore the dynamics of retraction, we develop agent-based models where actors on networks share and spread beliefs. Our models make the following, fundamental assumption: there is an asymmetry in the way a new finding spreads versus a retracted one. While individuals are apt to share novel information, they only tend to share retractions when the topic of conversation already centres on the false information. This assumption is consistent with Grice's maxim of relation that, in conversation, one should be relevant (Grice, 1975).

In our models, false beliefs can take a long time to eliminate after retraction. Additionally,

[1] Reuben pled guilty on 22 February 2010, and was sentenced to six months imprisonment; see Massachusetts (2010).

[2] See also Budd et al. (1998); Neale et al. (2010).

we find that whenever information gets old, in the sense that individuals stop sharing information after some time frame, false beliefs can persist indefinitely even when a retraction is issued. This occurs without any biased reasoning—we assume that any individual exposed to a retraction will change their mind. The persistence of false beliefs is a direct result of the fact that some individuals who received false beliefs from their neighbours happen never to receive a retraction.

Additionally, we consider the conditions under which retractions are more or less successful. We find there can be unexpected interactions between how long a retraction is delayed and how effective it is. In particular, a retraction that is introduced later—i.e., once a false belief is held widely—may be more efficacious, because it is relevant to a greater number of individuals. We also explore how network structure influences these processes. We examine small-world networks and preferential-attachment networks to see what effect the location of a retraction has on its success and find that retractions are more successful when issued from the original source. Additionally, we show that homophily—i.e., disproportionate in-group communication—can slow the spread of a retraction, especially if it is introduced in a subgroup where the false belief is not widely-held.

Our paper proceeds as follows. In section 2, we give an overview of relevant literature and introduce the modelling framework upon which this paper draws. In section 3, we present the simplest model explored here and describe its fundamental behaviour. Sections 3.2, 3.3, and 3.4 extend these results to several different scenarios intended to tease out the dynamics of retraction. Section 4 concludes.

## 2. Retraction and Contagion

As we have seen, retraction of a scientific paper does not always work the way it should. There are two things we should distinguish here. First is whether articles are cited after retraction, and second is whether scientists continue to hold beliefs now known to be false.

Our models actually consider the second question: how might false beliefs persist in the face of retraction/refutation of a result? However, the empirical literature focuses on citations, rather than underlying beliefs. Given scientific norms against citing known falsehoods, we take this literature to provide evidence (albeit imperfect) about false beliefs in the face of retraction.

Papers typically are retracted due to error, fraud, or failure to replicate (Wager and Williams, 2011; Steen, 2011; Fang et al., 2012), and retractions have become increasingly common (Cokol et al., 2008; Grieneisen and Zhang, 2012; Steen et al., 2013; Madlock-Brown and Eichmann, 2015). Importantly, study after study has found that even after retraction, papers continue to be cited, sometimes for years (Pfeifer and Snodgrass, 1990; Budd et al., 1998; Cor and Sood, 2018; Van Der Vet and Nijveen, 2016; Madlock-Brown and Eichmann, 2015). Some studies find declines in citation rate after retraction, others no change, or even an increase in citation rate (Cor and Sood, 2018; Van Der Vet and Nijveen, 2016; Madlock-Brown and Eichmann, 2015).

There is another, relevant kind of retraction which occurs in the news media. This often involves issuing an erratum or apology concerning a specific claim instead of retracting an entire article or report. There is little data about the effects of news-media retraction on belief. However, in many cases claims continue to spread widely after media retraction. During the COVID-19 pandemic, for instance, claims that the virus strips hemoglobin of iron were tweeted thousands of times after the Medium article originating them was removed. We do not claim here that retraction in science and retraction in media are just the same, but we think that many cases from both arenas can be captured by the models we develop.

Our models draw on the network-contagion modelling paradigm.[3] These models are widely used to represent the spread of both disease and belief (Hayhoe et al., 2017). The idea is that just as infections spread to susceptible individuals, in some cases beliefs spread

[3]Contagion models are a type of diffusion model (Rogers, 2012). See discussion in Levy and Nail (1993).

in social networks from those who hold them to those who do not.

There is an extensive literature on social-contagion models to which it would be impossible to do justice here. Nonetheless, the basic idea is simple. We use network models where the nodes are individuals (scientists, journalists, members of the public), and the vertices are communication channels between them. False information can spread from individual to individual. Information about a retraction can also spread; though, as we outline below, we assume there is an asymmetry in how false reports and retractions spread.[4]

Of course, not all beliefs are well-modelled as contagions. These models are best tuned to beliefs that spread unit-like from person to person and are easily adopted. In many cases, individuals depend on evidence to form beliefs in a rational or semi-rational way and can hold graded degrees of belief. This is especially true in scientific communities, where beliefs are expected to be evidence-based. (For this reason, philosophers of science have more often used the network-epistemology framework to represent the spread of scientific beliefs (Zollman, 2013; Goldman and O'Connor, 2019).)

What cases, then, are appropriate targets of the investigation here? The models apply to cases where beliefs are adopted relatively unreflectively. For instance, scientists may trust free-standing claims published by peers—e.g., that a particular hormone causes hunger or that a drug effectively reduces pain. Media consumers may trust journalists who claim there is a fire in some city, that *Moonstruck* is a great movie, or that SARS-CoV-2 strips iron from hemoglobin. In cases of beliefs that are controversial or that depend upon background theory, so that individuals engage in calculation or reasoning before adopting them, the models will be less applicable. Additionally, the models will apply best in cases where retractions/refutations are apparent—i.e., where it becomes uncontroversial that the original claim is, in fact, false.

[4]At least one previous paper—Hui et al. (2011)—has used a contagion-type network model to investigate the dynamics of retracted information. However, they assume their misinformed agents *exit* the network, rather than continuing to share information, which makes these models a poor fit to most real-world cases.

## 3. Model and Results

In this section, we present a base model and results. We also examine variations on this foundation to obtain a clearer picture of the dynamics of false and retracted information.

Suppose we have a population of $N$ individuals. Each has a belief about the world, contingent on information they have received. In particular, we consider *false* information, representing an erroneous or fraudulent scientific finding, or a misleading news item, and also *retracted* (or *corrected*) information. Individuals can thus have three belief states: neutral (before receiving any information), false (having received false information), and true or retracted (having received the retraction).

We start simulations with the majority of the population holding neutral beliefs. One individual holds a false belief. A retraction is introduced to one individual either at the start of the simulation or after some number of rounds. At each time step, we pair two network neighbours at random to interact. This is done by randomly selecting a focal individual and then randomly selecting a partner from their neighbours.[5]

Beliefs spread as follows. If one individual holds the false belief and the other a neutral belief, we assume the false belief always spreads. The idea is that the individuals pass on new information per the contagion framework discussed in Section 2. Retracted beliefs also spread in this way, but only to those who already hold the false belief. As mentioned in the introduction, this captures the idea that sharing some novel, albeit false, information is more apt than sharing a retraction of that same information, because a retraction is parasitic on the context. It is only interesting or relevant to those who already hold false beliefs. We might alternatively interpret the model in the following way: if one individual mentions a

[5]Many models of network contagion/diffusion assume a probabilistic flow of opinion that happens between each new adopter and all their neighbours—i.e., node $M$ adopts an opinion in round $t$, and with probability 0.3 all her neighbours will adopt that opinion in $t_1$. We instead assume probabilistic pairings and a deterministic flow of information when these pairings happen. We still take our model to fall under the umbrella of this framework since, unlike other epistemic-network models, we do not represent evidence or complex belief states.

false belief to a partner who knows it has been retracted, the partner will then share the retraction. (To be clear, there are cases not well captured by these models where retractions are of interest independent of initial findings.)

### 3.1. Complete Network

We begin with a population connected in a complete network, as in Figure 1. This means every agent is eligible for pairing with the focal individual on every given round. (We might also describe this as a population without a network, where individuals meet randomly for interaction.) When, after some time step, no further belief-revision can occur, we say that the

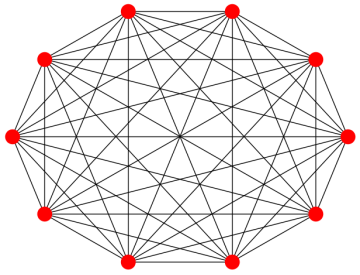


Figure 1: A complete network, with a population of $N = 10$ agents.

population is in a *stable state*. For a model of this sort, of any size, several population states are stable. For all of these, it is the case that no individual in the population holds a false belief. In other words, for the model to be in a stable state, every individual must hold either neutral or retracted information. This fact actually holds of a broader set of networks—those that are *connected* (meaning there is a path between every pair of nodes). Intuitively, this occurs because the retraction can spread to all individuals with the false belief, but not to individuals with a neutral belief. Since all individuals are connected, eventually each individual with a false belief will pair with someone holding the retraction and end up with the true belief. Either all members of the network are exposed to the false belief, and end up, eventually, with the true belief; or else, some of them never see the false belief, and at the end will have neutral beliefs that are stable because there are no false beliefs left in the

network.

**Proposition 1.** *For any connected network, a population configuration is stable if and only if no individual holds a false belief.*

*Proof.* See appendix A. □

Furthermore, the population will always converge toward one of these stable states given enough time. This fact is stated in Corollary 1.

**Corollary 1.** *For a connected network, in the limit, the population configuration always reaches a stable state.*

*Proof.* This follows from the proof of Proposition 1. □

Since the long-term fate of false belief is reasonably predictable, in the sense just outlined, for now, we are interested in the 'medium-run' results. How long does false information persist in this set-up? How widely does it spread? How do alterations influence this process? To answer these questions, we run simulations.[6]

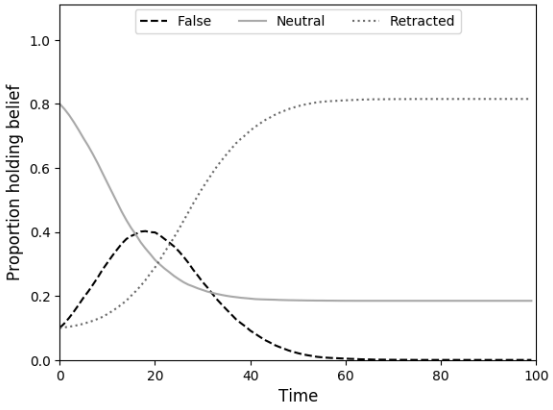
We examine populations of $N = 10, 50, 100, 1000$ individuals. In each case, one individual from the population is given false information, and one the retraction, at the outset. Simulations proceed round by round. Reported results are averages across simulations.[7] In each case, the population converges to a stable configuration (as expected). The typical behaviour of the model involves first the spread of the false belief, sometimes saturating the population (or coming close), and subsequently the spread of the retraction until all individuals hold

[6]The simulations are run using the Mesa agent-based modelling framework in Python3. See https://github.com/projectmesa. The code for our simulations is publicly available at [REMOVED FOR REVIEW].

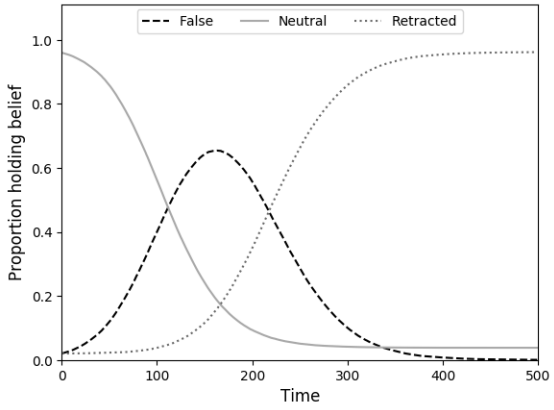
[7]We average the results of 1000, 1000, 1000, and 100 episodes, respectively, for networks of size $10, 50, 100,$ and 1000. In the largest population, simulations take longer to run. For this reason, we mostly focus on results from smaller populations where we can gather more data. Throughout, we ran each simulation long enough to reach a stable state

true or neutral beliefs. In figure 2, we can see this process for three population sizes. The $x$-axis tracks time (note the different time scales), and the $y$-axis tracks the proportion of the population in the three possible belief states.

(a) $N = 10$




(b) $N = 50$




(c) $N = 1000$

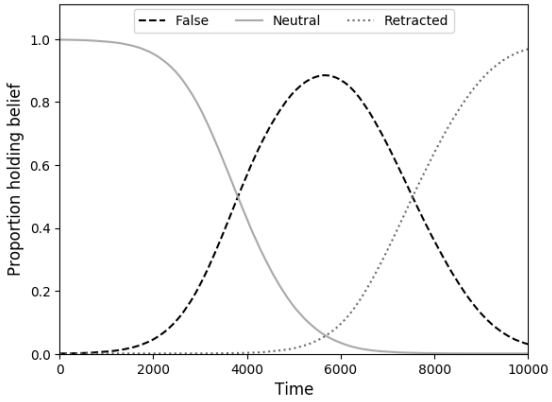



Figure 2: Simulation results for a population on a complete network. $N$ is the size of the population.

There are a few things to notice. A group of 10 individuals sees, on average, 40% of the population holding false beliefs at some point. When we increase the population to 1000 individuals, this becomes 90%. In other words, for a larger population, on average, false beliefs spread further. This is because, for smaller networks, the false belief is easier to nip

in the bud. Though the proportion of individuals holding retracted and false beliefs at the outset are the equivalent for a given population size (10% of the population for $N = 10$ and 0.1% of the population for $N = 1000$), for smaller networks, it is more likely that early pairings bring all those with the false belief in contact with the retraction.[8] For similar reasons, there appears to be a strong, positive relationship between the size of the population and the average length of time (number of time steps) for which an agent holds her false belief, as is evident in figure 3.


Figure 3: Length of time false beliefs are held for populations of different sizes. For larger populations, false beliefs are held longer on average.

These observations could be outlined analytically, given that pairings are based solely on probability distributions for the complete network—i.e., they depend entirely upon how many individuals hold each type of belief at a given time step and how large the population is. While the results here are perhaps unsurprising, we now extend this model by looking at some variations.

[8]By way of example, at the outset of a simulation with $N = 10$, the false belief is an order of magnitude more likely to spread than the retraction. In contrast, for $N = 1000$, the false belief is *three* orders of magnitude more likely to spread than the retraction.

### Delayed Retraction

We assume in the base model that the false information and the retraction enter the network at the outset of an episode. This assumption is perhaps accurate with something like real-time fact-checking during a political debate; however, in the case of replication and retraction of scientific studies, there is usually a (potentially significant) delay before the retraction enters into the population. For instance, Fang et al. (2012) find that it takes an average of three years for a finding to be retracted. In some cases, the discovery of fraud can trigger retraction of an author's older articles, including ones published many years previously.

We delay the introduction of retracted information by a parameter, DELAY. Otherwise, the simulations are run as before. The delay generally tends to increase the number of individuals who ever hold the false belief. It also increases the average amount of time individuals hold false beliefs. However, for relatively short delays, there is little impact on the average time false beliefs are held. This is shown in figure 4. As is apparent, for $N = 100$ there is virtually no change for DELAY $\leq 200$. This is because if a retraction is issued when relatively few individuals hold the false belief, it spreads very slowly, since few individuals will take it up. If it is issued once false beliefs have saturated the community, each interaction is one where the retraction will spread, so it catches on more quickly. As we will see later in the paper, this means that, perhaps surprisingly, under some conditions, later retractions are successful.

### Timed Novelty

We also previously assumed that agents share information indefinitely. However, in many cases new information is more readily shared than old. We model this as follows. Each agent—upon receiving novel (false or retracted) information—only shares it for a specified time frame. This small, realistic change means that the analytic results in Proposition 1 and Corollary 1 no longer hold. Now false information may stably persist. This happens when

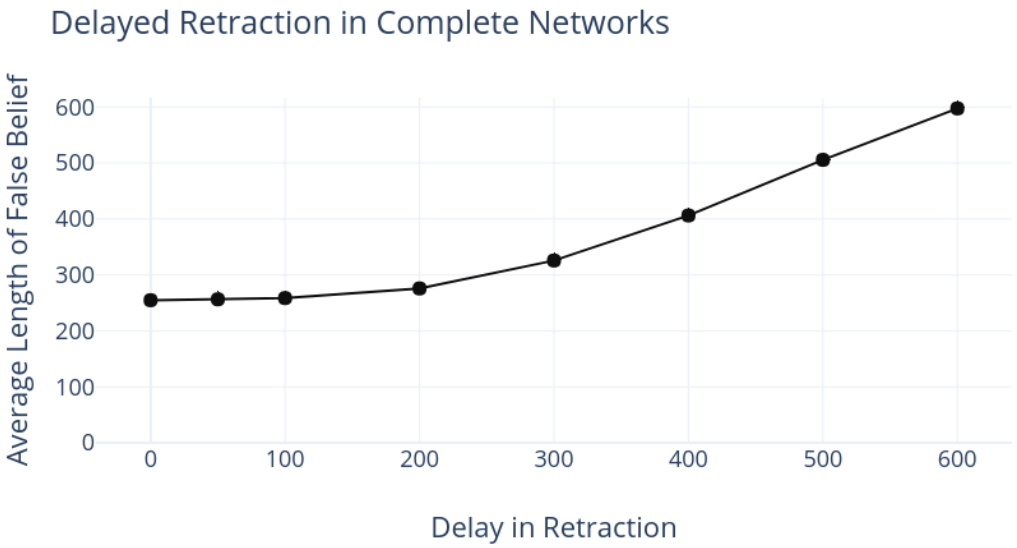



Figure 4: As retractions are delayed, false beliefs are held for a longer time on average. For short delays, there is relatively little effect. $N = 100$

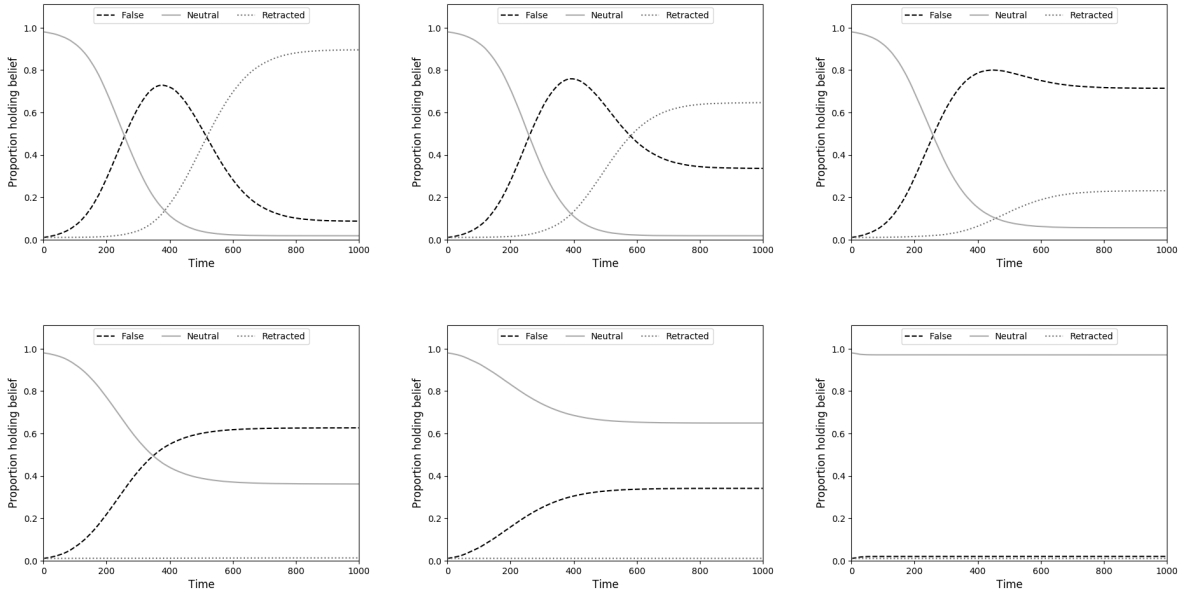
enough time has passed that no one shares the retraction, even though some individuals in the network hold false beliefs.

As before, in each case, we give one individual from the population false information and one individual the correction. Figure 5 shows results for population size $N = 100$.[9] The $x$-axis tracks time, and the $y$-axis tracks the average belief state of the population over simulations.

As is apparent, when the window for which individuals share beliefs is long, simulations are much like the previous models (Figure 5a). The false belief spreads and then is supplanted by the true belief. As the sharing window gets shorter, though, false beliefs start to persist alongside retracted beliefs because the retraction is no longer shared (Figures 5b, 5c). As the window grows shorter still, the retraction does not spread at all, and only false and neutral beliefs persist (Figures 5d, 5e). For the shortest time windows, neither belief spreads

[9]Results are qualitatively similar for $N = 50, 1000$. Because the population is so small when $N = 10$, the behaviour of the model is slightly different, but with timed novelty, false beliefs can persist indefinitely in this model as well.

(a) Time frame = 400 (b) Time frame = 300 (c) Time frame = 200




(d) Time frame = 100 (e) Time frame = 75 (f) Time frame = 25

Figure 5: Simulation results for the model with timed novelty on a complete network with a population of $N = 100$ individuals.

(Figure 5f). In other words, there is a regime of moderate sharing where false beliefs are common at the end of a simulation.

Why do we see these effects? For the false information to begin to spread in the first place, the individual holding the false belief must be chosen within the early rounds before they stop sharing. Since there are 100 individuals in the population, each equally likely to be picked on a given round, the agent holding the false information has a reasonably low probability of being chosen within a short frame of time. It is harder still for the retraction to spread since this requires that the individual with the retracted belief be selected in the first rounds, and also meet a neighbour with the false belief.

Since false beliefs are stable in these models we can ask: under different regimes, what proportion of the population ends up with a false belief? Figure 6 shows the qualitative trend visible in figure 5. Each colour band represents the average proportion of individuals at that

end state for some time frame of belief-sharing. As the time frame grows, neutral beliefs decrease, the proportion of stable false beliefs grows and then shrinks, and the proportion of retracted beliefs grows.[10]

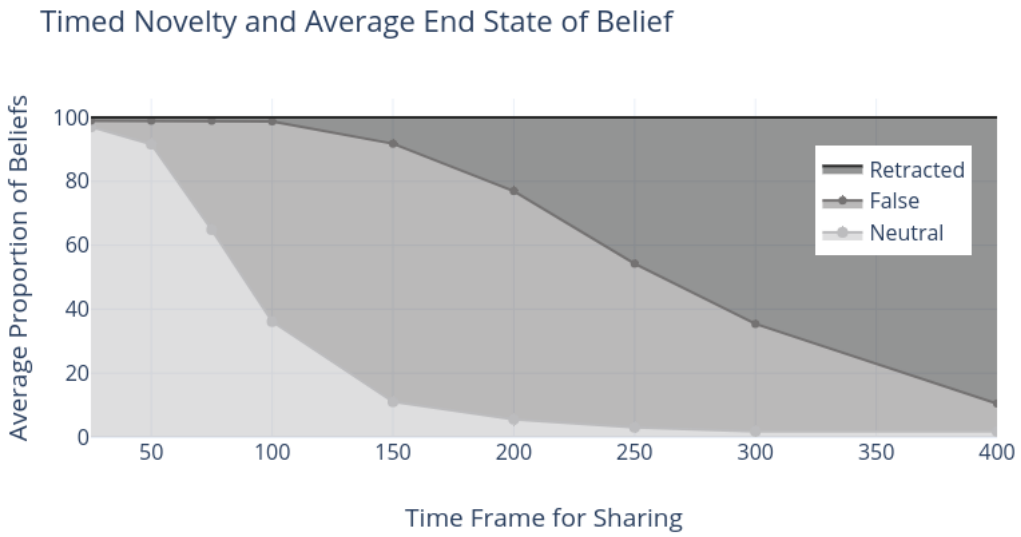



Figure 6: As sharing time increases, false beliefs become more prevalent and then less prevalent. $N = 100$, DELAY $= 0$

We have been discussing independent variations to the base model, but there are interaction effects between these parameters. In particular, a delayed retraction can interact with timed novelty in the following way. When there is only one individual with false information and one individual with retracted information, on any given round, there is a minuscule probability that the retracted information can spread. However, if almost everyone in the population already holds a false belief, this probability is much higher. We have already seen that a delay in retraction does not necessarily lead to a significant increase in average length of false belief, for this reason. However, in some cases, a delay in retraction can actually

[10]Reported results to this point have been averages over simulations. However, in these models, there is significant variation in the level of false belief at the end of runs for particular sets of parameter values. For some runs, the false belief will gain traction, while for others, it may never spread at all. For some runs, the retraction will spread widely due to an accident of history, and for others, the retraction does not reach many individuals. This means that some information is lost in the data we have presented so far. This is especially true for longer time frames.

*improve* belief because a population where the false belief is widely held will be more receptive to the retraction. It can catch on like a contagion. Figure 7 shows this. As the delay increases, we see a decrease in the final number of false beliefs because the delay makes the retraction relevant during the time frame where individuals are sharing it.

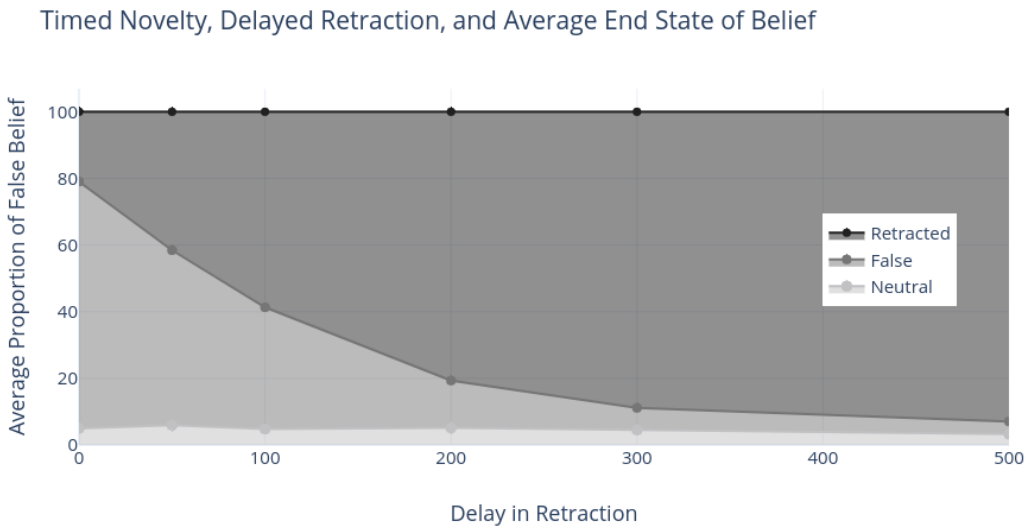



Figure 7: For a fixed sharing time frame of 200, as the delay in retraction increases, final false beliefs decrease, and retracted beliefs increase. $N = 100$

### 3.2. Small-World Networks

To this point, we have considered a trivial network structure which assumes every individual meets every other randomly. But some scientists are regular communicators, and others do not interact at all. We can vary network structure to restrict the individuals with whom any particular agent can share information—two agents are eligible for pairing just in case they are connected.

We consider small-world networks because many real-world social networks exhibit small-world properties (Telesford et al., 2011). Small-world networks are defined by short average path-length—i.e., the distance between any pair of nodes is relatively short—and high clustering coefficients. The clustering coefficient is a measure of how dense the connections are

for individual nodes: If your friends are all friends with one another, then you have a high clustering coefficient. Small-world networks also tend to have hubs, which are nodes with higher-than-average connections—e.g., a popular individual at the centre of a clique is a 'hub' for social interactions.[11]

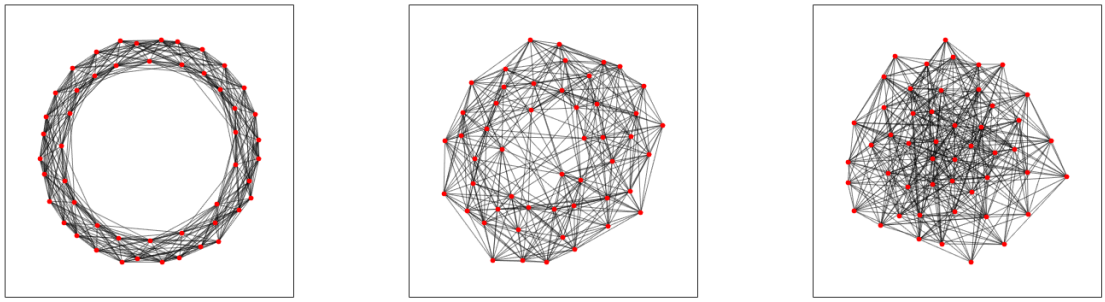
We generate networks according to the *Watts-Strogatz* model (Watts and Strogatz, 1998). The algorithm begins with a network with $N$ nodes, each of which connects to its $K$ nearest neighbours. Then, for every node, $n_i$, it takes every edge connecting $n_i$ to its $K/2$ rightmost neighbours and re-wires it with probability $p$. Rewiring is done such that the new link connects $(n_i, n_k)$, where $k$ is chosen at random, subject to the constraints that there are no loops, and no duplications. We examine models with $(k, p) = (8, 0.1), (16, 0.07), (32, 0.016)$.[12] Figure 8 shows an example of a regular lattice, a small-world network, and a random network. As the figure makes clear, the small-world network is related to both of these. The more rewiring in the algorithm, the further the network is from a lattice, and the closer to the random network. Small worlds exist for intermediate values between these extremes.

With a non-trivial network structure, we must now specify the relationship between the source of the false belief and the source of the retraction. They might originate at the same node. This captures a situation in which, e.g., a specific individual (lab, journal, institute) publishes a result and then subsequently retracts it. Alternatively, a retraction might originate at a different node if another journal or lab publishes a failure to replicate, or if one news outlet invalidates another's claim.

The qualitative results from our base model hold for small worlds. We discuss simulations for $N = 100$. When individuals never stop sharing information, the only stable states are

[11]Note that small-world networks generally display an *over*-abundance of hubs compared to real-world networks. The preferential-attachment networks considered in the next section do not.

[12]The parameters for $k$ were chosen somewhat arbitrarily, but they correspond to an individual, on average, knowing between 1/10 and 1/3 of the population. Once we decided upon these values for $k$, we empirically tested a variety of values for $p$ and calculated the 'small-worldness' of the resultant network using the $\omega$ measure described in Telesford et al. (2011). The values we chose consistently achieved $\omega$ close to zero.



(a) lattice (b) small world (c) random

Figure 8: Particular instances of (a) a lattice network, (b) a small-world network, and (c) a random network. In each case, the population consists of 50 individuals, and the average degree for each node is 8.

ones in which there are no false beliefs. Delaying the introduction of retracted information again tends to increase the average amount of time individuals hold false beliefs. As before, when we introduce timed novelty to our model, stable false beliefs are possible. And as before, delay can create a surprising benefit in such cases.

Does the location of a retraction influence its success? We looked at simulations where the retraction was either introduced 1) by the same agent who introduced the false belief or 2) by another, random agent in the network. We find that introducing the retraction in a different spot leads to more, stable false belief (assuming there is some time frame for sharing). Figure 9 shows this for a particular model. This happens because the retraction, when introduced in the same location, can chase and overtake false beliefs.[13] When introduced in another location, it takes longer for the retraction and false beliefs to come into contact, and thus the false belief is harder to eradicate. As we will discuss in the conclusion, this may have important policy implications for journals.

[13]Notice that for the model where the same agent issues the retraction, increasing delay first slightly increases false belief and then decreases it. This is because when the retraction is issued right away, there may be no time for the false belief to spread at all. So, there are few false beliefs. But when the delay continues to increase, it improves the uptake of the retraction.

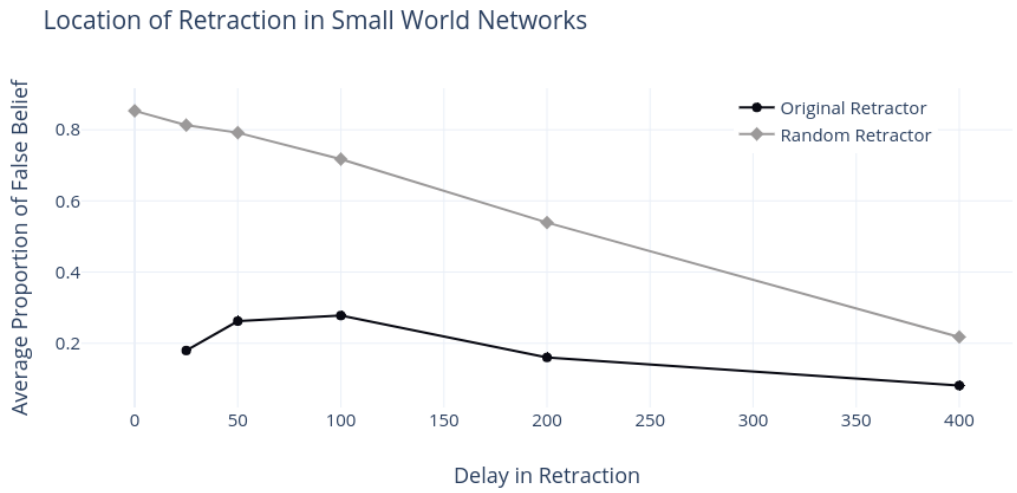



Figure 9: False beliefs are less prevalent when the same source issues a retraction, $N = 100, K = 8$, time frame for sharing $= 200$.

### 3.3. Preferential-Attachment Networks

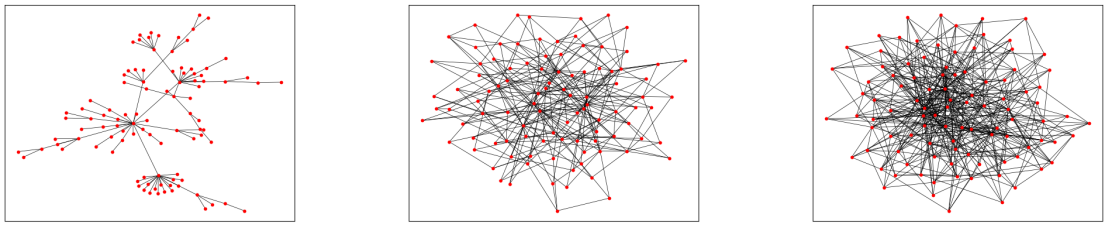
As noted, small-world networks accord with some empirically observed properties of social networks. However, unlike small-world networks, many real-world networks have been observed to have *scale-free* properties such that some small number of individuals are very highly connected, while most are not.[14] In particular, scale-free properties are common in citation networks, scientific collaboration networks, and on the internet (Barabási and Albert, 1999; Barabási et al., 2002; Albert and Barabási, 2002; Steyvers and Tenenbaum, 2005).[15] This might correspond, for instance, to scientific communities where some individuals are highly connected, and others are marginally so.

We now look at networks generated according to the Barabási–Albert preferential-attachment model (Barabási and Albert, 1999). The algorithm begins with a connected network of $m$ individuals. New nodes (up to $N$) are added one at a time. Each connects to $m$ existing nodes with probability directly proportional to the number of links the nodes already have. There-

[14] In a scale-free network, the asymptotic degree distribution follows a power law.

[15] Though the scale-free nature of some networks has been contested; see Clauset et al. (2007); Milojević (2010); Broido and Clauset (2019).

fore, existing nodes with many links tend to receive more new attachments, e.g., new articles are proportionally more likely to cite 'famous' (already heavily cited) articles rather than newer articles with fewer citations. Figure 10 shows an example of networks for $m = 1, 3, 5$. We examine networks with $m = 1, 2, 3, 4, 5$.



(a) $m = 1$ (b) $m = 3$ (c) $m = 5$

Figure 10: Example of preferential-attachment networks.

Our results are generally robust in these new networks. Timed novelty can lead to the indefinite persistence of false beliefs, and delayed retraction can sometimes improve the saturation of a retraction. As with small-world networks, for preferential-attachment networks, we can ask: how does the location of retraction influence outcomes? We were particularly interested in cases where a founding, central node introduced a false result. We then looked at treatments where either 1) the retraction was introduced by this same node, or 2) the last node introduced the retraction. This might correspond to situations where either a well-established scientist retracts their own finding, or where a relative newcomer to the community refutes them.

We found that, across parameter values, retraction is more effective when introduced by the original, founding node. It is not entirely surprising. There are two advantages to a central node issuing a retraction. First, as with the small-world networks, there is an advantage of a retraction coming from the same place because it can chase the same paths the false belief took. Second, there is a benefit when the retraction is issued by a highly connected node, with relatively short paths to the rest of the network. Figure 11 shows

these results for different values of $m$. Note that when $m$ is low, there is relatively little false belief because the sparse network structure impairs its spread. In all cases, there is a clear benefit to retraction from the original node.

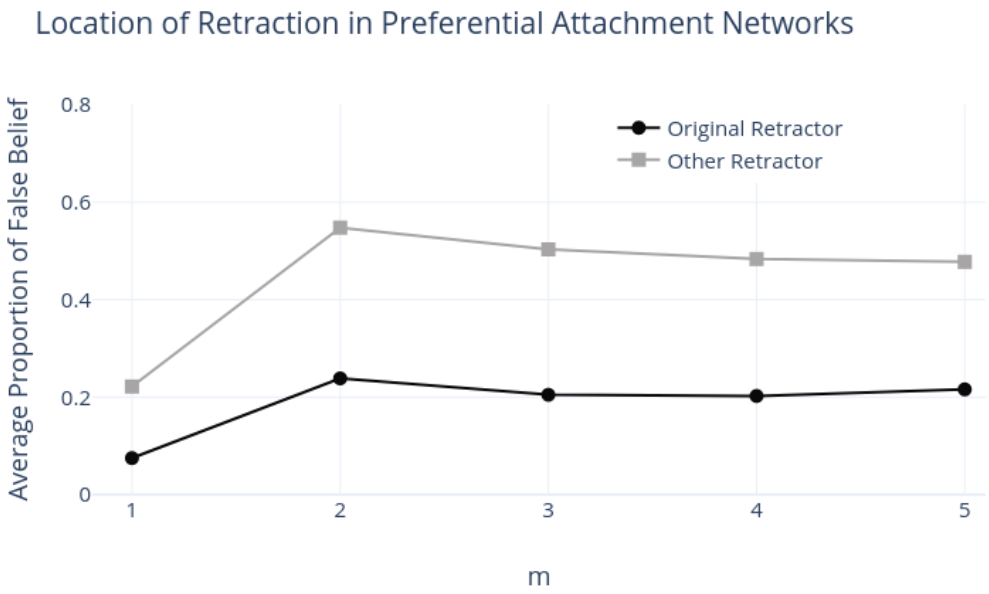



Figure 11: False beliefs are less prevalent when a central node issues a retraction of its own falsehood, rather than when a relatively less central node does. $N = 100$, Delay=100, time frame for sharing = 200

### 3.4. Homophilic Networks

Sometimes populations fracture into political camps or research communities that take different approaches. Such networks are especially relevant to thinking about failures of retraction, since retractions accepted by one subgroup may be ignored by another.

To consider this sort of case, we look at how *homophily* affects the persistence of retracted information. Network homophily describes the tendency for nodes that are 'similar' in some respect (i.e., in-groups) to be more likely to attach to one another than 'dissimilar' (i.e., out-group) nodes. Because of this tendency, in homophilic networks, we see highly connected subgroups with relatively fewer connections between. Figure 12 shows an example.

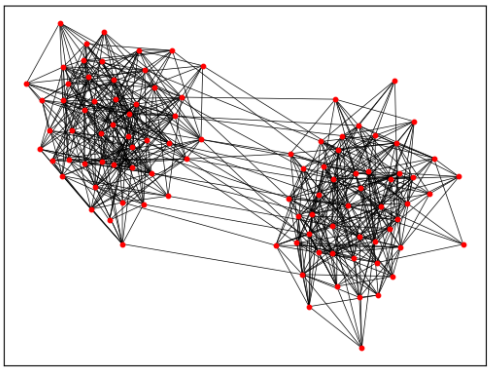


Figure 12: Example of a network exhibiting homophily. Each subpopulation has $N = 50$ agents. The probability that a particular individual has a connection with a member of her in-group is significantly higher than the probability that she has a connection with a member of her out-group. In this case, $p_{in} = 0.25$ and $p_{out} = 0.10$, respectively.

There are two reasons we examine homophilic networks. First, we see subgroups of this sort in many real-world networks, including scientific communities.[16] Additionally, social networks surrounding political orientation are often homophilic (Himelboim et al., 2016). Second, by splitting the population into subgroups, we can examine the effect that distinct sources of information have on the persistence of false information in the population as a whole and within each subpopulation.

We always assume that for $N$ individuals, each subpopulation consists of $N/2$ individuals. For each individual, $n_i$, in the network, there is some probability, $p_{in}$, that $n_i$ is connected with a given member of her in-group, and there is some other probability, $p_{out}$, that $n_i$ is paired with a member of her out-group. In Figure 12, for example, $p_{in} = 0.25$ and $p_{out} = 0.10$.

What happens in models with homophily? General results are similar to those reported. Compared to the model with a complete network, retracted beliefs tend to spread less quickly

[16]For example, among medical researchers in the United States, there is a sharp divide between those who believe that 'chronic Lyme disease'—i.e., a form of the disease that resists short-term antibiotic treatment—is a fiction and those who report efficacy of long-term antibiotics to relieve symptoms (O'Connor and Weatherall, 2019).

in homophilic networks, meaning that on average neutral and false beliefs persist longer, and in models with timed novelty fewer individuals ever reach retracted beliefs.

In cases with a high level of homophily between the groups, we find that retractions are less successful when introduced in the group that did not generate the original false belief. This is unsurprising since homophily means that it takes longer for false beliefs to reach the other group, making retraction less relevant and more likely to stop spreading.[17] Once the retraction does manage to spread, there are fewer links by which it can travel to the group that originated the false belief. Figure 13 shows this.

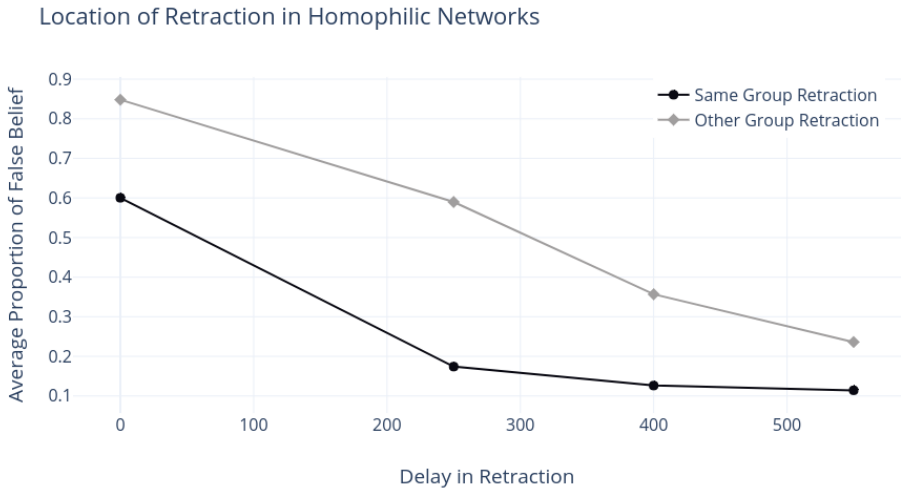



Figure 13: Homophilic networks may have more persistent false belief when retraction is introduced to a different partition from where it originated. $p_{in} = 0.4, p_{out} = 0.004$, $N = 100$, timed novelty stops after 200 rounds

Additionally, both the location of the original false belief and the location of the retraction can influence the relative levels of false belief in the two subgroups. For simplicity, let us call the groups 1 and 2. We assume the false belief is always introduced in group 1, and the retraction can be introduced in either. In most cases, group 1 tends to hold false beliefs for more extended periods than 2, since the false belief originated in their area of the network.

[17]Perhaps more surprising is that we see little effect for lower levels of homophily. We are not sure why only extreme homophily values exhibit these tendencies.

Though in cases where both the false belief and the retraction show up in group 1, it will sometimes be the case that the false belief, but not the retraction, manages to spread effectively to group 2. In such cases, more members of group 1 will hold the false belief at some time or another, but they'll also learn the retraction faster. When the retraction is introduced in group 2, though, group 1 holds false beliefs for longer in all models. If the individuals with the incorrect belief do not receive a correction in their own subgroup, they are left in a state of false belief as the retraction slowly trickles back to them. Figure 14 shows data supporting these claims. We report average end beliefs for groups 1 and 2, for cases where the retraction is introduced in 1 (same) and 2 (other). When the retraction is in the same subgroup, that group ends up with many more retracted beliefs; and, for these parameter values, shorter average false beliefs. When the retraction is introduced in group 2, group 1 has dramatically higher levels of false belief and lower levels of retracted beliefs.

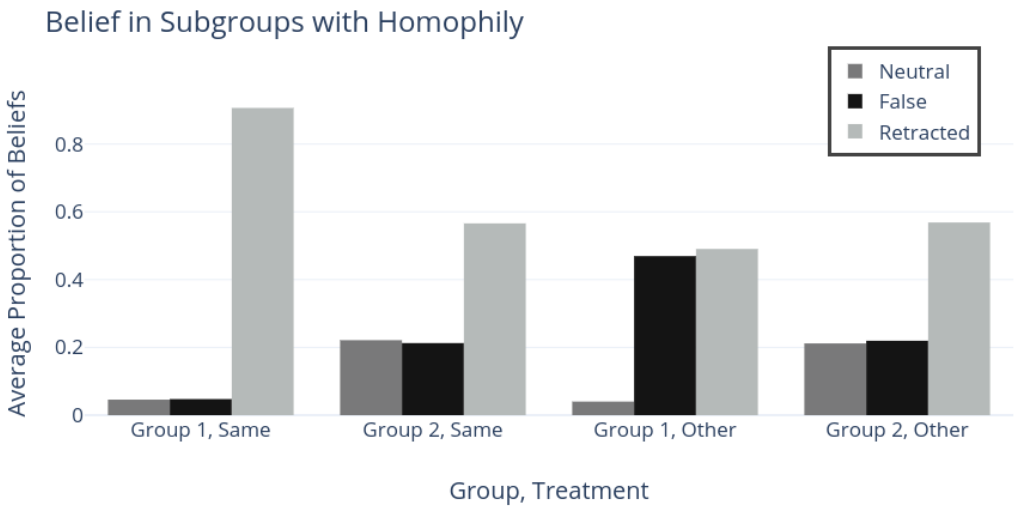



Figure 14: Average length of belief for two subgroups in a homophilic network. Results are for group 1 and group 2, for cases where the retraction is introduced in the same subgroup (1) and the other subgroup (2). $p_{in} = 0.4, p_{out} = 0.004$, $N = 100$, Delay=400, timed novelty stops after 200 rounds

## 4. Conclusion

To summarise our findings: we find that contagion-type models are a useful tool in exploring the dynamics of retraction. They show why false beliefs can persist indefinitely, even in light of a retraction, when agents stop sharing new beliefs. They illustrate how delays might influence the success of a retraction, including the surprising finding that in some cases delay might improve the saturation of a retraction. They show how network structure can impact the success of retraction, and, in particular, how retractions are less successful when they do not come from the original source. In homophilic networks, retractions are more successful when they originate in the same group as the false claim.

Simplified models like the ones we present here must always be treated carefully when applied to real-world cases. In particular, there are mismatches between model and target in our work that may attenuate the relevance of the results. We will now talk about a few of these mismatches before outlining in greater detail what we think these models can do.

One of these gaps is that we do not model prominent, or central communicating agents such as a journal, or academic search engine, that spread ideas to large portions of a population at once. One might worry that in science authors will always check with these sorts of central agents before adopting a new belief, or citing a source, thus invalidating our models. Empirically, though, we know that this does not always happen. Instead, authors often pull citations directly from citing papers, rather than looking to the original source (Broadus, 1983). Moreover, as we made clear earlier in the paper, the existence of these central agents does not seem to stop the widespread citation of retracted work. Still, this kind of structure should impact the flow of information in scientific communities, and below we will discuss this further.

Another issue relates to the representation of agents' cognition. Our agents have only three belief states—neutral, false, or retracted. However, in many cases, results of scientific studies are not easily deemed to be true or false but are controversial. For this reason, as

noted above, the models will not apply to a wide range of cases where beliefs vary in degree. Additionally, the agents in our populations have none of the special reasoning biases humans exhibit. Thus, there are no instances of, e.g., confirmation bias, where individuals seek out all and only 'facts' that support their beliefs. They do not yield to conformist bias—i.e., imitating the beliefs of neighbours for social reasons. Further work might look at whether the results here hold up under more complex representations of agents.

Our models also assume that agents are not motivated to shape the beliefs of those around them. Whenever agents learn that a piece of information is false, they stop sharing it and start sharing the retraction. However, it seems scientists often continue to cite and share their own retracted findings. Madlock-Brown and Eichmann (2015) show this is a common practice and that scientists who do it boost their post-retraction citation count. One reason for this may be that scientists often seek *credit*—attention and good reputation from those in the community (Merton, 1973).

Scientific journals and research institutions are also incentivised to maintain both readership and reputation, which may make them both reluctant to retract papers and to communicate these retractions to readers (Unger and Couzin, 2006; Wager and Williams, 2011; Madlock-Brown and Eichmann, 2015). For instance, there seem to be many cases in which journal retractions are overly vague, or imply an error, when, in fact, they were the result of discovered fraud (Wager and Williams, 2011; Fang et al., 2012).[18] Many philosophers of science have used Merton's framework to model scientists as taking part in a 'credit economy' (Kitcher, 1990; Bright, 2017; Heesen, 2018). A next version of our model might include credit motivations as well as epistemic motivations for our agents. In other cases, retracted information is shared by those with political or economic motivations. We might also consider agents who seek to shape the landscape of belief.[19]

[18]News sources may do the same thing. Writing about news retractions, Craig Silverman reports that news sources often try to downplay the fact that they were wrong (McWilliams, 2013).

[19]Philosophers of science have successfully used network-epistemology models to investigate situations

Despite the limitations of these models, there are several appropriate takeaways. First, they are useful for hypothesis-generation and directing further empirical research. This is especially true for cases where empirical work is lacking, such as in the case of uptake of media retractions. To give an example: our results suggest that moderate delays in retraction may sometimes make them more effective. Although this is not an obvious hypothesis to test prior to this modelling work, it is worth examining given the theoretical support generated here. It is also worth considering how the location of a retraction in real networks influences outcomes. Does the source matter, especially in homophilic groups, as we suggest?

Second, the models here provide tools for thinking about how to improve current systems to make retractions more effective. (Improvements that can be empirically tested.) What solutions do they suggest? We cannot easily alter the network connections between human individuals, laboratories, or research institutes. Additionally, we probably cannot convince agents to keep talking about retractions for a longer period than they usually would. However, we might be able to institute changes for central communicators like those mentioned above—journals and academic search engines. Imagine the addition of a node into our network models that communicates with a large proportion of the population and continues to share information about the retraction actively and indefinitely. Such a node would move the population towards something like the networks we discussed first, where falsehood is always eventually stamped out. If each individual is connected to a node that continues to share the retraction, they should eventually get that information.

This also implies that real organisations should be more active about communicating retractions. For instance, in searching Google scholar, it is easy to yield retracted research papers as results without also seeing the retraction. A better practice would involve tying these search results together. Journals should implement editorial policies that check to see whether cited sources have been retracted, and then ask authors to remove these sources

like this (Holman and Bruner, 2015, 2017; Weatherall et al., 2018; Lewandowsky et al., 2019).

where appropriate. This would, in effect, be a policy designed to promote continued and widespread sharing of the retraction.

Our results also suggest it is important for retractions to be spread by the same sources that originated a false belief. If a refutation is published in another field or subfield, it might not be effective. The journal that originally published the false result should publicize this refutation to their own readers. Likewise, if a news source proves that another is wrong, it is important that the original news source share this information as well. Journalists who, for instance, publicise false claims should try to use the same venues to clarify matters.

In sum, while these models are highly simplified, they are useful in directing further research into retraction. In particular, they help yield hypotheses regarding network structure and community design that may help improve retraction and the elimination of false beliefs.

## Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1535139 and under Grant No.1922424. Many thanks for their support. Thanks to the UC Irvine Social Dynamics seminar and to members of the Department of Philosophy at Australian National University for feedback on this work.

## References

Albert, Réka and Albert-László Barabási (2002). "Statistical mechanics of complex networks." *Reviews of Modern Physics*, *74*(1), 47–97.

Barabási, Albert-László and Réka Albert (1999). "Emergence of scaling in random networks." *science*, *286*(5439), 509–512.

Barabási, Albert-László, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek (2002). "Evolution of the social network of scientific collaborations." *Physica A: Statistical mechanics and its applications*, *311*(3-4), 590–614.

Bornemann-Cimenti, Helmar, Istvan S. Szilagyi, and Andreas Sandner-Kiesling (2016). "Perpetuation of Retracted Publications Using the Example of the Scott S. Reuben Case: Incidences, Reasons and Possible Improvements." *Science and Engineering Ethics*, *22*(4), 1063–1072.

Bright, Liam Kofi (2017). "Decision theoretic model of the productivity gap." *Erkenntnis*, *82*(2), 421–442.

Broadus, Robert N (1983). "An investigation of the validity of bibliographic citations." *Journal of the American Society for Information Science*, *34*(2), 132–135.

Broido, Anna and Aaron Clauset (2019). "Scale-free networks are rare." *Nature Communications*, *10*(4), 1017.

Buckwalter, Joseph A., Vernon T. Tolo, and Regis J. O'Keefe (2015). "How Do You Know It Is True? Integrity in Research and Publications: AOA Critical Issues." *The Journal of Bone and Joint Surgery. American volume*, *97*(1), e2.

Budd, John M, MaryEllen Sievert, and Tom R Schultz (1998). "Phenomena of retraction: reasons for retraction and citations to the publications." *Jama*, *280*(3), 296–297.

Clauset, Aaron, Cosma Rohilla Shalizi, and M. E. J Newman (2007). "Power-law distributions in empirical data." *SIAM Review*, *51*(4), 661–703.

Cokol, Murat, Fatih Ozbay, and Raul Rodriguez-Esteban (2008). "Retraction rates are on the rise." *EMBO reports*, *9*(1), 2–2.

Cor, Ken and Gaurav Sood (2018). "Propagation of Error: Approving Citations to Problematic Research.".

Fang, Ferric C, R Grant Steen, and Arturo Casadevall (2012). "Misconduct accounts for the majority of retracted scientific publications." *Proceedings of the National Academy of Sciences*, *109*(42), 17028–17033.

Goldman, Alvin and Cailin O'Connor (2019). "Social Epistemology." *The Stanford Encyclopedia of Philosophy*. Ed. Edward N. Zalta. Fall 2019 edition. Metaphysics Research Lab, Stanford University.

Grice, H. Paul (1975). "Logic and Conversation." *Syntax and Semantics, Vol. 3: Speech Acts*. Ed. Peter Cole and Jerry L. Morgan. New York: Academic Press, 41–58.

Grieneisen, Michael L and Minghua Zhang (2012). "A comprehensive survey of retracted articles from the scholarly literature." *PloS one*, *7*(10), e44118.

Hayhoe, Mikhail, Fady Alajaji, and Bahman Gharesifard (2017). "A Polya Contagion Model for Networks." *IEEE Transactions on Control of Network Systems*.

Heesen, Remco (2018). "Why the reward structure of science makes reproducibility problems inevitable." *The Journal of Philosophy*, *115*(12), 661–674.

Himelboim, Itai, Kaye D Sweetser, Spencer F Tinkham, Kristen Cameron, Matthew Danelo, and Kate West (2016). "Valence-based homophily on Twitter: Network analysis of emotions and political talk in the 2012 presidential election." *new media & society*, *18*(7), 1382–1400.

Holman, Bennett and Justin Bruner (2017). "Experimentation by industrial selection." *Philosophy of Science*, *84*(5), 1008–1019.

Holman, Bennett and Justin P Bruner (2015). "The problem of intransigently biased agents." *Philosophy of Science*, *82*(5), 956–968.

Hui, Cindy, Malik Magdon-Ismail, Mark Goldberg, and William A Wallace (2011). "Effectiveness of information retraction." *2011 IEEE Network Science Workshop*. IEEE, 133–137.

Kazil, Jackie, Nathan Vērzemnieks, et al. (2014). "Project Mesa: Agent-based Modeling in Python 3+.".

Kitcher, Philip (1990). "The division of cognitive labor." *The journal of philosophy*, *87*(1), 5–22.

Levy, D. A. and P. R. Nail (1993). "Contagion: A Theoretical and Empirical Review and Reconceptualization." *Genetic, Social, And General Psychology Monographs*, *119*, 233–284.

Lewandowsky, Stephan, Toby D Pilditch, Jens K Madsen, Naomi Oreskes, and James S Risbey (2019). "Influence and seepage: An evidence-resistant minority can affect public opinion and scientific belief formation." *Cognition*, *188*, 124–139.

Madlock-Brown, Charisse R and David Eichmann (2015). "The (lack of) impact of retraction on citation networks." *Science and Engineering Ethics*, *21*(1), 127–137.

Massachusetts, United States Attorney's Office (2010). "Anesthesiologist Sentenced on Health Care Fraud Charge." `https://www.justice.gov/archive/usao/ma/news/2010/June/ReubenScottSentencingPR.html`. *Offices of the United States Attorneys, Press Release*, *Jun*, 2010.

McWilliams, James (2013). "Journalism is Never Perfect: The Politics of Story Corrections and Retractions." 20 December 2013. `https://psmag.com/social-justice/journalism-never-perfect-politics-story-corrections-retractions-71700`. *Pacific Standard*.

Merton, Robert K (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Milojević, Staša (2010). "Modes of collaboration in modern science: Beyond power laws and preferential attachment." *Journal of the American Society for Information Science and Technology*, *61*(7), 1410–1423.

Neale, Anne Victoria, Rhonda K Dailey, and Judith Abrams (2010). "Analysis of citations to biomedical articles affected by scientific misconduct." *Science and engineering ethics*, *16*(2), 251–261.

O'Connor, Cailin and James Owen Weatherall (2019). *The Misinformation Age: How False Beliefs Spread*. New Haven: Yale University Press.

Pfeifer, Mark P and Gwendolyn L Snodgrass (1990). "The continued use of retracted, invalid scientific literature." *Jama*, *263*(10), 1420–1423.

Rogers, Everett M. (2012). *Diffusion of Innovations*. 5 edition. New York: Simon and Schuster.

Shafer, S. L. (2015). "Tattered Threads." *Anesthesia and Analgesia*, *108*(5), 1361–1363.

Steen, R Grant (2011). "Retractions in the scientific literature: is the incidence of research fraud increasing?." *Journal of medical ethics*, *37*(4), 249–253.

Steen, R Grant, Arturo Casadevall, and Ferric C Fang (2013). "Why has the number of scientific retractions increased?." *PloS one*, *8*(7), e68397.

Steyvers, Mark and Joshua B. Tenenbaum (2005). "The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth." *Cognitive Science*, *29*(1), 41–78.

Telesford, Qawi K., Karen E. Joyce, Satoru Hayasaka, Jonathan H. Burdette, and Paul J. Laurienti (2011). "The Ubiquity of Small-World Networks." *Brain Connectivity*, *1*(5), 367–375.

Unger, Katherine and Jennifer Couzin (2006). "Even retracted papers endure." *Science*, *312*(5770), 40–41.

Van Der Vet, Paul E and Harm Nijveen (2016). "Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature." *Research integrity and peer review*, *1*(1), 3.

Wager, Elizabeth and Peter Williams (2011). "Why and how do journals retract articles? An analysis of Medline retractions 1988–2008." *Journal of medical ethics*, *37*(9), 567–570.

Watts, Duncan J. and Steven H. Strogatz (1998). "Collective Dynamics of 'Small-World' Networks." *Nature*, *393*, 440–442.

Weatherall, James Owen, Cailin O'Connor, and Justin Bruner (2018). "How to Beat Science and Influence People." DOI: 10.1093/bjps/axy062. *British Journal for Philosophy of Science*.

Zollman, Kevin JS (2013). "Network epistemology: Communication in epistemic communities." *Philosophy Compass*, *8*(1), 15–27.

## Appendix A. Proofs

**Proposition 1:** *For any connected network, a population configuration is stable if and only if no individual holds a false belief.*

*Proof.* ($\Leftarrow$) Assume that no individual in the population holds a false belief. Then every individual in the population either has neutral information or retracted information.

On a particular trial, we pair two individuals, $A$ and $B$. Either $A$ and $B$ have the same information, or they have different information. If they have the same information—both neutral or both retracted—then they do not update, so the state-configuration remains unchanged. This leaves the case where they have different information. Without loss of generality, assume that $A$ holds a neutral belief, and $B$ holds a retracted belief. In this case, *ex hypothesi*, they do not share information. Therefore, the state configuration remains unchanged.

($\Rightarrow$) We proceed via the contrapositive.

Assume at least one individual in the population holds false information. By assumption, at least one individual in the network holds a retracted belief. Because the network is connected, there is at least one path of vertices connecting these individuals. On every round, there is a positive probability that any two neighbours meet. Thus, from any starting time step, in the limit, all neighbours on this path will be selected to meet with probability 1. Consider moving down the path from the individual with false belief. If the individual with false belief has a neighbour with neutral or retracted belief, the configuration is thus not stable. If their neighbour has a false belief, but if the next neighbour has a neutral or retracted belief, the configuration is not stable, etc. Since the final node on the path holds the retracted belief, the sub-network consisting in this path is not stable, and thus the entire network is not in a stable configuration.

□