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*
University of Copenhagen

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.

1. Introduction

Scott S. Reuben, an American anaesthesiologist and Professor of Anaesthesiology and Pain Medicine, was a prolific and influential researcher in pain management. Over the course of more than a decade, he 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 2009, Reuben was convicted of fabricating data in 21 of these major publications. All of these articles were subsequently retracted (Shafer, 2015). Before retraction, these articles had obtained nearly 1200 citations. By 2014—five years after his articles were retracted—a case study found that roughly half of these retracted articles continued

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

Draft of October 7, 2019 *PLEASE CITE THE PUBLISHED VERSION IF AVAILABLE!*

to be cited consistently; however, only 1/4 of the citing articles clearly stated *that* Reuben's work had been retracted (Bornemann-Cimenti et al., 2016).[1]

It seems remarkable that a scientific finding of this sort would be so widely used in the literature after it had been officially withdrawn as fraudulent; but this is by no means an isolated case.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 are interested in examining the perpetuation of false information even in light of a correction. When retracted scientific information continues to influence members of scientific communities and members of the public, it can cause harm. White et al. (2009), for instance, point out that the retracted Reuben articles, "compromise every meta-analysis, editorial, and systematic review of analgesic trials that included these fabricated findings and every lecture and Continuing Medical Education course on perioperative analgesia that included these studies" (1367). It should be clear why threatening the integrity of research on anaesthesia, in particular, has direct adverse effects. Furthermore, outside the scientific realm, popular media sources regularly retract claims, but, in many cases, the original claim will reach members of the public who never see the retraction and so remain uninformed. Poorly informed voters and decision-makers can cause harms just as poorly informed doctors and scientists do.

To explore the dynamics of retraction, we introduce agent-based models built upon the social-contagion/diffusion framework, 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. The intuition underlying this assumption is

[1]The scale of this fraudulent activity had profound consequences for research in the area. Also, Reuben's research was supposed to address "topical questions for which anaesthetists, surgeons, and patients seek answers" (Moore et al., 2010, 329). In other words, the work had direct consequences for patient treatment.

[2]See also Budd et al. (1998), who find 94% of post-retraction citations they investigated treat the work as valid. In the biomedical field Neale et al. (2010) find that only 5% of post-retraction citations mention that the article in question is retracted.

that individuals are apt to share information that is novel but only tend to share retractions when the topic of conversation already centres around the false information. This assumption is consistent with Grice's maxim of relation that, in conversation, one should be relevant (Grice, 1975). Intuitively, a retraction is more likely to be a non-sequitur than a novel bit of information itself.

In all the models we explore, false beliefs can take a long time to eliminate after retraction. Additionally, we find that whenever information gets old, in the sense that individuals stop volunteering to share information after some time-frame, false beliefs can persist indefinitely in a social network even when a retraction is issued. This phenomenon 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 demonstrate that 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 more significant number of individuals. We also explore how network structure influences these processes. In particular, we look at small-world networks, which mimic many real-world human networks, to see what effect the location of a retraction has on its success. As we show, retractions tend to be more successful when they originate from the same source as an original false belief. Additionally, we consider whether the presence of homophilous structure—i.e., disproportionate in-group communication—prevents the uptake of a retraction. As we show, high levels of homophily can slow the spread of a retraction, especially if it is introduced in a subgroup that does not widely hold the false belief.

This exploration is particularly important in light of the replication crisis. In social psychology, oncology, and the broader social sciences, a significant number of key studies have

failed to replicate, and they are sometimes subject to retraction as a result.[3] Additionally, Ioannidis (2005); Prasad et al. (2013) find that a large percentage of studied medical practices are eventually reversed—i.e., contradicted by future studies. When study findings are overturned, it is crucial to understand why individuals might fail to learn that the findings are no longer current.

Our paper proceeds as follows. In section 2, we give an overview of relevant literature. We discuss the ways that false scientific beliefs sometimes persist in light of retraction. We also introduce the modelling framework this paper draws upon, discussing social contagion and how it has traditionally been modelled. In section 3, we present the simplest model explored here and describe the fundamental behaviour of this model. Sections 3.2 and 3.3 extend these results to several different scenarios intended to tease out the dynamics of retraction. In section 4, we conclude.

2. Retraction and Contagion

2.1. Retraction

As we saw in the introduction, retraction of a scientific paper does not always work the way that it ought to. There are two things we should be careful to distinguish here. First is whether and when articles continue to be cited after retraction, and second is whether individuals continue to hold beliefs introduced or supported by retracted findings that are now known to be false. As will become apparent, our models consider the second question: how might individual false beliefs persist in the face of retraction/refutation of a result? However, it is usually easier to investigate citation patterns in studying retraction, which means that much of the empirical literature focuses on citation, rather than underlying beliefs. We take it that this empirical literature does provide evidence (albeit imperfect)

[3]See, for example, Begley and Ellis (2012); Collaboration et al. (2015); Camerer et al. (2016).

about the prevalence of false beliefs in the face of retraction given scientific norms against citing known falsehoods.

Moreover, note that actual retraction of a paper in a scientific journal is not synonymous with admitted falsehood of results. Sometimes retractions occur because authors self-plagiarise, fail to obtain permission to use patient data, fail to obtain co-author permission to publish, or are discovered to have a conflict of interest (Madlock-Brown and Eichmann, 2015); or, sometimes only part of an investigation is fraudulent or contains an error. In other cases, results are thoroughly refuted—i.e., shown to be entirely invalid—but are not actually retracted. Again, this means that data on retraction in scientific journals will not perfectly match what we are modelling—the persistence of false scientific beliefs in the face of a retraction or clear refutation—but will still generally be useful. Note that throughout the paper, we use the term "retraction" to describe what we are modelling. This should be taken as a convenient shorthand for cases where previous findings or claims have been clearly overturned.

Papers typically are retracted due to error, fraud, or failure to replicate. Previous authors have found that error is more common than fraud in the biomedical fields (Wager and Williams, 2011; Steen, 2011). Though, Fang et al. (2012) find that fraud, broadly construed, is the more common cause, and they argue that misleading retraction announcements have led to an under-diagnosis of fraud. In recent years, retractions for all reasons have become more common (Cokol et al., 2008; Grieneisen and Zhang, 2012; Steen et al., 2013; Madlock-Brown and Eichmann, 2015). Study after study has found that even after retraction, papers continue to be cited, often at very high rates (Pfeifer and Snodgrass, 1990; Budd et al., 1998; Cor and Sood, 2018; Van Der Vet and Nijveen, 2016; Madlock-Brown and Eichmann, 2015).[4] Some find declines in citation rate after retraction, others no change, or even an increase

[4]Fang et al. (2012) note that in 2012 the twenty most highly-cited retracted articles shared thousands of citations.

in citation rate (Cor and Sood, 2018; Van Der Vet and Nijveen, 2016; Madlock-Brown and Eichmann, 2015). This also does not seem to be a problem that clears up quickly; instead, papers often continue to be cited for years after a retraction has been issued (Cor and Sood, 2018).

As noted, there are crucial issues with the continued citation of retracted papers. In cases where relevant findings have been invalidated, continued citations spread false scientific beliefs to new readers. Also, as Cor and Sood (2018) point out, using false claims in the development of new scientific work can promote new errors. Begley and Ellis (2012), discussing cancer research, point out that, "Some non-reproducible preclinical papers had spawned an entire field, with hundreds of secondary publications that expanded on elements of the original observation, but did not actually seek to confirm or falsify its fundamental basis" (532). This is worrying for obvious reasons.

Additionally, there is a different kind of retraction to which the models we present in this paper can speak: the retraction of claims made by the news media. Retraction of this sort is common. Unlike scientific retraction, it often involves issuing an erratum or apology concerning a specific claim, rather than an entire article or report.[5] However, there is little data about the effects of news media retraction on the beliefs of consumers of this media. For this reason, the models discussed here may be especially helpful in informing our understanding of what happens to incorrect media information in the light of subsequent retraction.

2.2. Contagion/Diffusion in Networks

The underlying idea behind social contagion is that socio-cultural artefacts (such as affect, attitudes, beliefs, and behaviours) can spread through populations as if they were infectious. In such cases, mere contact, as opposed to, e.g., rational deliberation, may be sufficient for the

[5]Though sometimes not. In 2002, a Fox news anchor falsely reported that PETA had dressed Ohioan deer in orange hunter vests. The network retracted the whole story (Hudson, 2012).

spread of such artefacts. Contagion as an explanation for social phenomena was popularised in Baldwin (1894); Le Bon (1896); De Tarde (1903), and has received a significant amount of attention in the social sciences from the mid-20th century on.[6]

This sort of contagion process is often modelled as an *epidemic* in a network. As Hayhoe et al. (2017) point out, a contagion model might represent the spread of a disease, but it may also represent the spread of a computer virus, a rumour, or competing opinions in a social network.[7] Typically, epidemics in these models are assumed to spread via pairwise interactions between 'infected' individuals and healthy individuals—and through a single exposure. There is an extensive literature on social contagion models. As such, it would be impossible to do any serious justice to the literature here.

Nonetheless, the basic idea is quite simple. We use network models where the nodes are individuals (scientists, journalists, members of the public), and the vertices are channels of communication 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.[8]

At least one previous paper has used a contagion type network model to investigate the dynamics of retracted information. Hui et al. (2011) consider agents who can receive graded messages pushing in two directions, and who combine these messages in deciding how to act. Their model, though, assumes that if agents decide to act, they do so by *exiting* the network, rather than continuing to share information with neighbours. For this reason, their model is a poor fit to the cases we investigate here.

Of course, not all beliefs are well-modelled by contagion models. In particular, these mod-

[6]See the discussion in Levy and Nail (1993).

[7]See, for example, Garetto et al. (2003); Adar and Adamic (2005); Rogers (2012); Kim et al. (2014).

[8]Contagion models are a type of diffusion model (Rogers, 2012). Two other types of diffusion models, which we will not consider here, are social influence models—where individuals in the population only adopt new beliefs when enough of their neighbours have—and social learning models—where individuals assume beliefs only after they have seen empirical evidence that supports such a belief. See Young (2009).

els are 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 (Bala and Goyal, 1998; Zollman, 2007).[9] We agree that the network epistemology model is a useful tool in studying epistemic communities. We also think, though, that sometimes information, ideas, and knowledge spread in a contagion-like manner even in scientific communities.

What cases, then, are appropriate targets of the investigation here? The models will apply to cases where information and ideas are picked up from social sources 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, or that a politician delivered a certain stump speech. In cases of beliefs that are controversial, or are closely related to background theory so that individuals engage in calculation or reasoning before deciding to adopt them, the models will be less applicable. Additionally, the models will apply best in cases where retractions/refutations are very clear—i.e., where it is uncontroversial that the original claim is, in fact, false. In cases where there is debate, or doubt, network epistemology type models that allow for representations of evidence and reasoning will be more appropriate.

[9]For more on this framework in philosophy of science see Mayo-Wilson et al. (2011); Holman and Bruner (2015); Rosenstock et al. (2017); Weatherall et al. (2018); Frey and Šešelja (2018). Zollman (2013) provides a review.

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 in a network of individuals.

Suppose we have a population consisting of N individuals. Each individual has a belief about the world, contingent on information they have received. In particular, there are two sorts of information that spread in our networks. The first is *false* information, representing an erroneous or fraudulent scientific finding, or a misleading news item. The second is *retracted* (or *corrected*) information. Individuals can thus have three belief states: neutral (before they have received 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 rounds. As noted, the N members of our population live on a network where each node is an individual, and each vertex is a social connection. At each time step, we pair two network neighbours at random to interact. This pairing is done by first randomly selecting a focal individual and then randomly selecting their interactive partner from their network neighbours.

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 here is that the individuals are passing on new information à la the contagion framework discussed in Section 2.2. Retracted beliefs also spread in this way, but we assume that retractions spread only to those who already hold the false belief—i.e., a retraction will not spread from an individual with a retracted belief to one with a neutral 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. New,

false information can spread without context, whereas the retraction 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 false belief to a partner who knows it has been retracted, the partner will then share the retraction.

3.1. Complete Network

In our first set of models, let us assume that the population is 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.) For a model of this sort, of any size, several

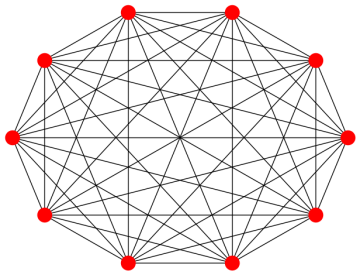


Figure 1: A complete network, with a population of $N = 10$ agents.

population states are stable in the sense that information is no longer transferred. For all of these, though, it is the case that no individual in the population holds a false belief. In other words, to be in a stable state, every individual must have neutral information or the retraction. 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 pairs with someone holding the retraction, and ends 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 the complete network, a population configuration is stable if and only if no individual holds a false belief.*

Proof. See Appendix A. □

Furthermore, we know that since any two individuals are eligible for random pairing on any given trial with a complete network, the population will always converge toward one of these stable states. This fact is stated in Corollary 1.

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

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

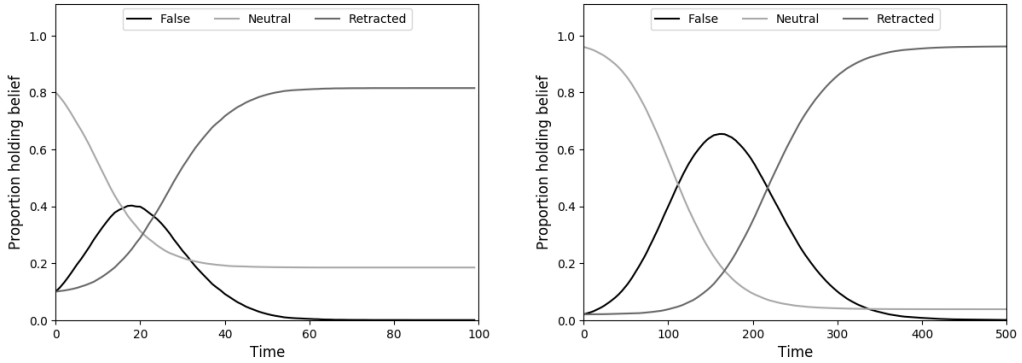
Since the long-term fate of false belief is reasonably predictable, in the sense just outlined, for now, we are primarily interested in the 'medium-run' results. How long does false information persist in this set-up? How widely does the false belief spread? How do alterations affect the persistence of false information? To answer these questions, we run simulations.[10]

We examine simulation results of this model with a population of $N = 10, 50, 100$, and 1000 individuals. In each case, one individual from the population is given false information at the outset, and one individual from the population is given the corrected information at the outset. Simulations proceed round by round as described. Reported results are averages

[10]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].

across simulations.[11] In each case, the population converges to a stable configuration. 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 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) $N = 1000$

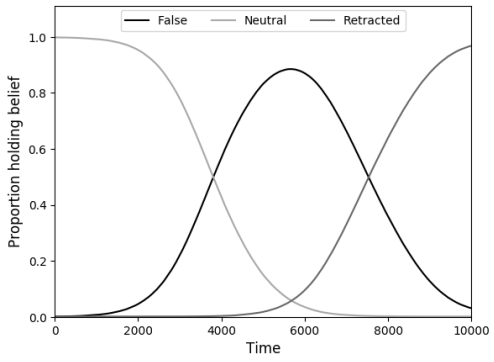


Figure 2: Simulation results for a population on a complete network. N is the size of the population.

[11] We average the results of 1000, 1000, 1000, and 100 episodes, respectively, for networks of size $10, 50, 100$ and 1000. This is because, 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. Additionally, we ran each simulation long enough to reach a stable state, though this length depended on the parameter details of the model.

There are a few things to notice here. 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, 90% of the population holds false beliefs at some point. 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. The single individual holding the retracted belief makes up 10% of the population when $N = 10$ and only 0.1% of the population when $N = 1000$. For smaller networks, it is more likely that early pairings bring all those with the false belief in contact with the retraction. For similar reasons, there appears to be a strong 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. The larger the population size, the longer the time each individual holds the false belief.

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.

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 about 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. (Recall from the introduction that, in the case of Reuben, some articles took more than 15 years to be found fraudulent, and were subsequently retracted.)

We now delay the introduction of retracted information into the population by a parameter, DELAY. The simulations are run as before. After DELAY time-steps, one individual in the population receives the retracted information, and the simulation continues. The delay in retraction generally tends to increase the number of individuals who ever hold the false belief. It also increases the average amount of time that individuals hold false beliefs in the network. However, for relatively short delays, there is little impact on the average time false beliefs are held. This is shown in figure 4, which outlines the average time individuals hold false beliefs as the delay increases. As is apparent, for $N = 100$ there is virtually no change for DELAY ≤ 200. This occurs because of the asymmetries in how false beliefs and retractions spread. If a retraction is issued when relatively few individuals hold the false belief, it spreads very slowly, since few individuals 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, paradoxically, under some conditions, it may be better to issue a retraction later.

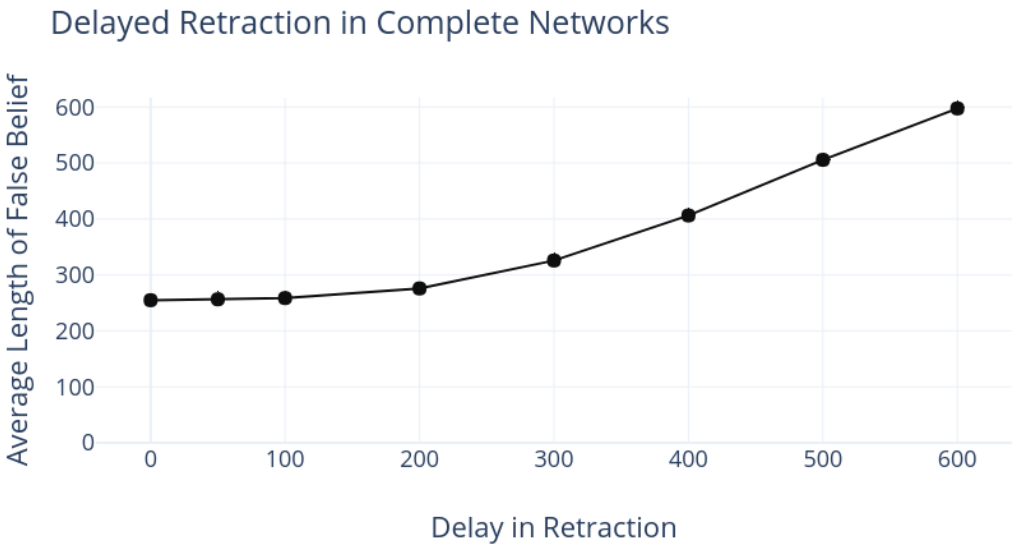


Figure 4: As retractions are delayed, false beliefs are held for a longer time on average. For short delays, there is relatively little effect. $N = 100$

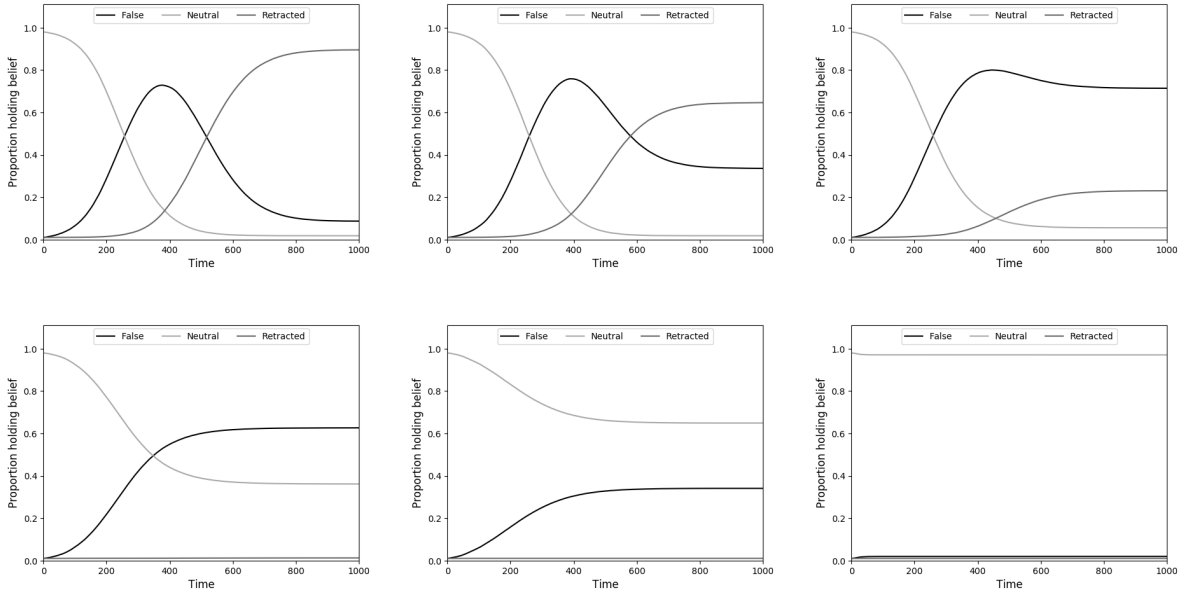
Timed Novelty

We also assumed in the previous simulations that agents share their information indefinitely. However, let us suppose that novel information is more readily shared than old information. We model this as follows. Each agent—upon receiving novel (false or retracted) information—only shares that information 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 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 corrected information at the outset. Figure 5 shows results for population size $N = 100$.[12] The x-axis tracks time, and the y-axis tracks the average belief state of the population over simulations.

[12]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.

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 (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.

In these models, false beliefs are stable, so it makes less sense to analyse them by considering the average length of time that individuals hold a false belief. Instead, we can ask: under different regimes, what proportion of the population ends up with a stable 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.

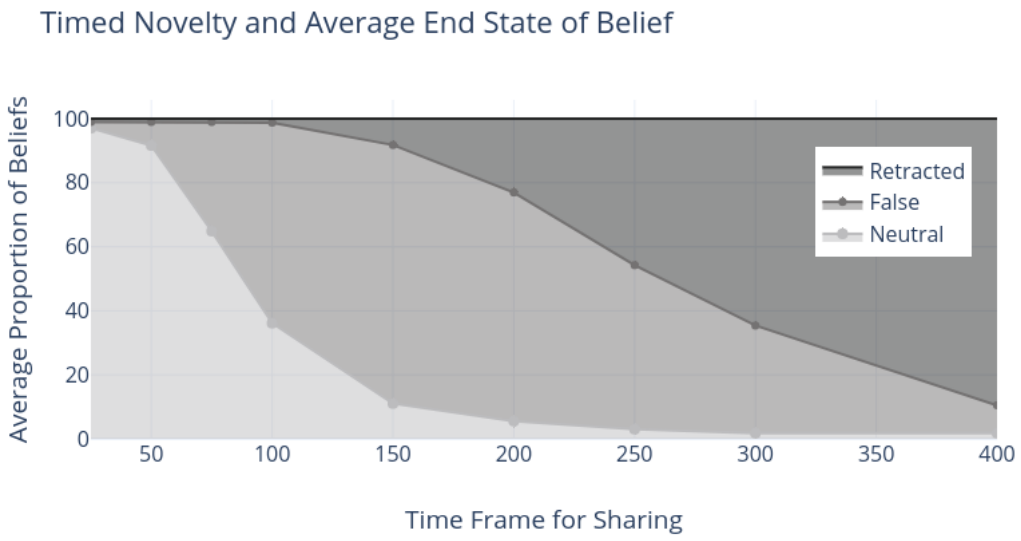


Figure 6: As sharing time increases, false beliefs become more prevalent and then less prevalent. $N = 100$, DELAY $= 0$

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. In particular, for longer time-frames, it tends to be the case that the retraction either never spreads, or reaches almost everyone. This means that the distribution of false beliefs is bi-modal—either they are widespread, or nearly non-existent. For shorter sharing time-frames, the range of levels of stable false belief is broader, including intermediate levels.

Note that we have been discussing independent variations to the base model, but there may be 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, then the probability that the retraction will begin to spread in the first place is much higher. We have already seen that a delay in retraction does not necessarily lead to a significant increase in the length of time false beliefs are held on average, for this reason. However, in some cases, a delay in retraction can actually *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 for sharing time-frame of 200. As the delay increases, we see a decrease in the final number of false beliefs and an increase in the number of retracted beliefs, because the delay makes the retraction relevant during the time frame where individuals are sharing it.

3.2. Small-World Networks

To this point, we have considered a trivial network where every individual meets every other randomly. This assumption is obviously not empirically accurate. Some scientists are regular communicators, and others do not interact at all. We can relax the assumption, though, by varying the underlying network structure of the model. In this way, we restrict the individuals

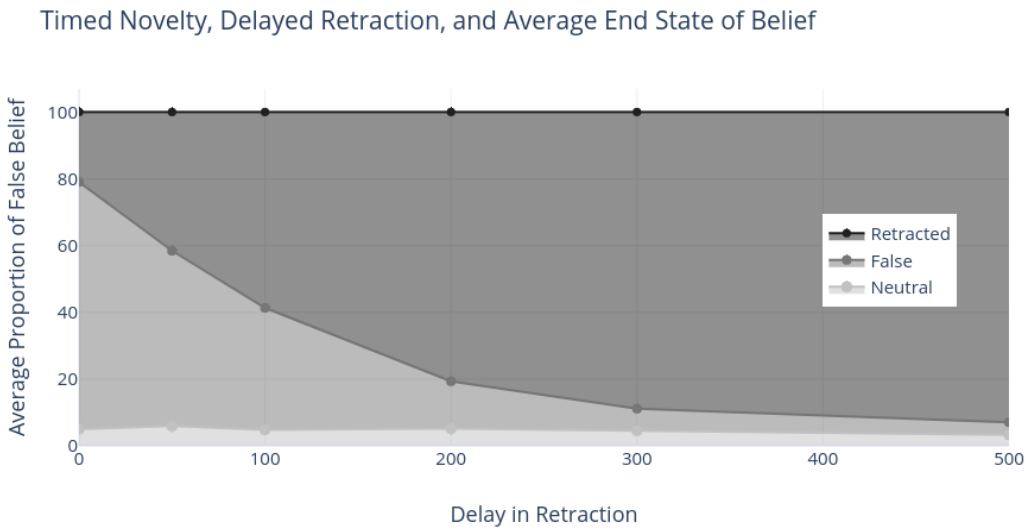


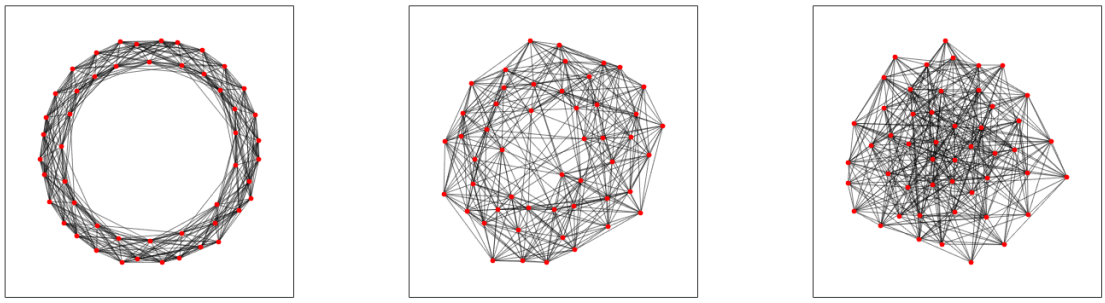
Figure 7: For a fixed sharing time-frame of 200, as the delay in retraction increases, final false beliefs decrease and retracted beliefs increase.

with whom any particular agent can share information—two agents are eligible for pairing just in case they are connected in the network.

We chose to consider small-world networks because many real-world human networks have been shown to 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 small—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; whereas, if your friends do not know one another, you have a low 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.

We generated networks according to the *Watts-Strogatz* model, which guarantees that the networks satisfy the small-world properties. 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 (1) $k \neq i$, and (2) $k \neq k'$, where the edge connecting $(n_i, n_{k'})$ already exists. (That is, no loops, and no duplication.) We specifically examine models with $(k, p) = (8, 0.1), (16, 0.07), (32, 0.016)$.[13] Figure 8 shows an example of a regular lattice, a random network, and a small-world 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.



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

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

Now that we have varied the network structure, we can again look at how factors like delayed retraction and timed novelty influence outcomes. 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. For instance, we can suppose they originate at the same node. This can capture a situation in which, e.g., a specific individual (lab, journal, institute) publishes a

[13]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 ω measure described in Telesford et al. (2011). The values we chose consistently achieved ω close to zero.

result and then subsequently publishes a retraction of that result. Alternatively, a retraction might originate at a different node if another journal or lab publishes a failure to replicate, or if one news outlet contradicts another. For this section, we will investigate models where the retraction is issued by the same source as the false belief, though in the next section, we will expand to different sources.

In general, the qualitative results from our base model still hold when we impose this network structure on our population. We discuss simulations for $N = 100$.[14] As with the base-model, when individuals never stop sharing the information that they have, the only stable states are ones in which there are no false beliefs. Delaying the introduction of retracted information into the population by the parameter, DELAY, again generally tends to increase the average amount of time that individuals hold false beliefs in the network. (However, with the structure of a small-world network, relatively short delays do have an impact on the average time false beliefs are held.) As before, when we introduce timed novelty to our model, the stability results no longer hold. It is now possible that false beliefs will persist alongside retracted beliefs.

One thing the network structure allows us to ask is: does the location of a retraction influence its success? To answer this, 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. 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 in the network.[15] When introduced in another location, it takes longer for the retraction and

[14] As was mentioned above, focusing on smaller populations allowed us to gather more data. Thus, we do not compare the results for $N = 1000$ between the base model (complete network) and the Watts-Strogatz model (small-world network).

[15] Notice that for the model where the retraction is issued by the same agent, increasing delay first slightly increases false belief and then decreases it. This is because when the retraction is issued too soon, 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.

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.

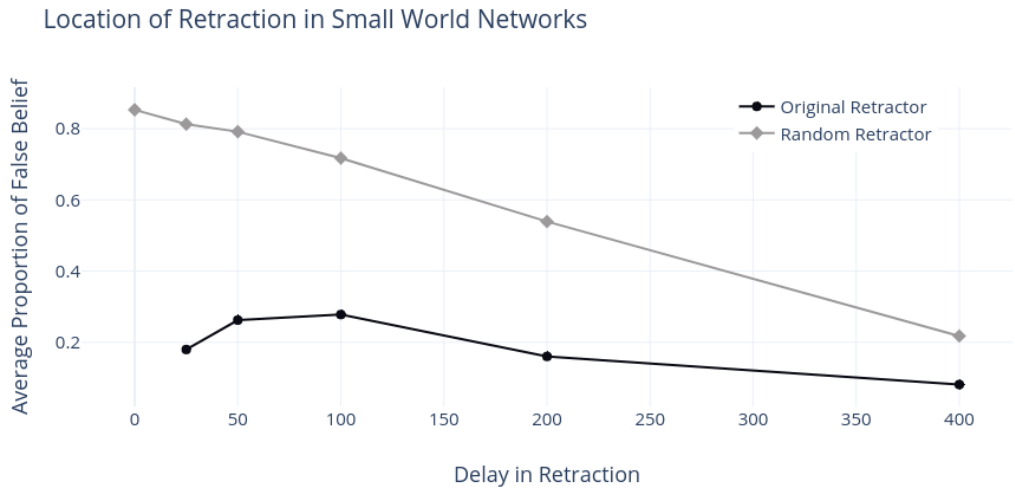


Figure 9: False beliefs are less prevalent when the same source issues a retraction, $N = 100, K = 8$, time frame for sharing $= 200$.

3.3. Homophilous Networks

The previous models have assumed a relatively homogeneous population. However, we might split the population into two subpopulations that could represent, for instance, two political camps, or two research communities that take different approaches. To do this, we need to 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 homophilous networks, we see subgroups that are highly connected within the group, and we see relatively fewer connections between groups. Figure 10 shows an example.

There are two reasons we examine homophilous networks. The first is that in many real-world cases, including in the scientific community, we see subgroups of this sort. For

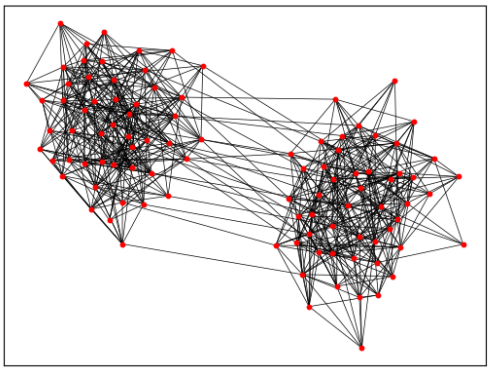


Figure 10: Example of a network exhibiting homophily. Each subpopulation has $N = 50$ agents. The probability that a particular individual has a connection with a member of her in-group is significantly higher than the probability that she has a connection with a member of her out-group. In this case, $p_{in} = 0.25$ and $p_{out} = 0.10$, respectively.

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 and persists in the body causing damage—is a fiction and those who treat hundreds of patients for this disease and report efficacy of long-term antibiotics to relieve symptoms. Doctors in these two communities are highly mistrustful of each other, and often avoid those in the other group (O'Connor and Weatherall, 2019). Additionally, social networks surrounding political orientation are often homophilous (Himelboim et al., 2016). In thinking about retracted news items, then, it makes sense to consider homophilous networks.

Second, by splitting the population into subgroups, as noted, 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 can ask, for instance, what happens when a retraction is issued within the same subgroup that originated a belief? What happens when it arises in a different subgroup?

We generated networks with homophily in the following way. 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 10, for example, $p_{in} = 0.25$ and $p_{out} = 0.10$, so that any given individual is 2.5 times more likely to be paired with an individual in her own subpopulation. Hence, we have two well-connected neighbourhoods with sparse connections between them.

What happens in models with homophily? When there is no timed novelty, and the false and retracted beliefs are randomly introduced, the general dynamics of the model are similar to those in the base model with complete network structure. The false belief spreads and is eventually replaced by the retracted belief. The longer the delay, the longer false beliefs are held, on average. As in our other models, timed novelty means that false beliefs can persist indefinitely. Moreover, under this regime, it is sometimes better to delay retraction so that it is more relevant when it is introduced. Compared to the model with a complete network, retracted beliefs tend to spread less quickly in homophilic networks, meaning that on average neutral beliefs 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.[16] 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 11 shows this. With no delay, there is little difference because the retraction fails to spread in either case. When a delay is added, though, there is a significant difference between the average proportion of false beliefs at the end of simulation

[16]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.

dependent on whether the retraction is introduced in the subgroup that originated the false belief, or in the other subgroup.

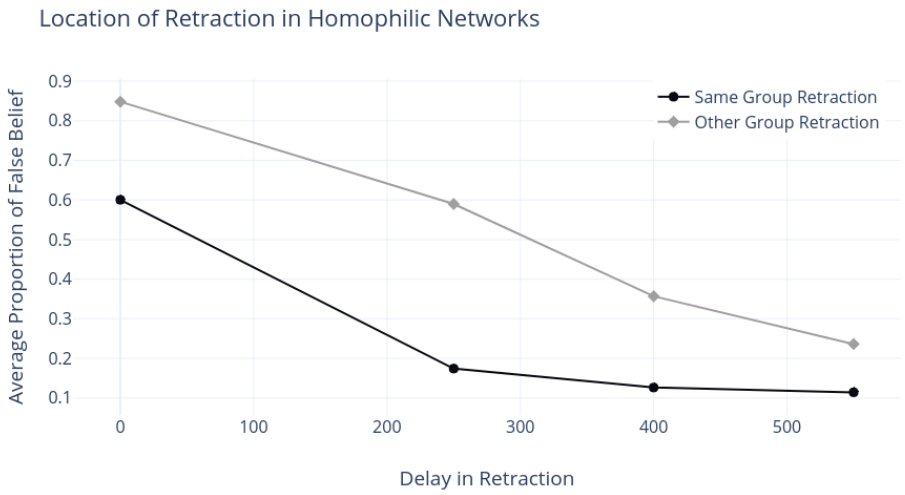


Figure 11: Homophilic networks may have more persistent false belief when retraction is introduced to a different partition from where it originated. $p_{in} = 0.4, p_{out} = 0.004$, $N = 100$, timed novelty stops after 200 rounds

Additionally, both the location of where the original false belief was introduced 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 will assume that 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. Though in cases where both the false belief and the retraction show up in group 1, it will sometime be the case that the false belief, but not the retraction manage 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 12 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 in 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.

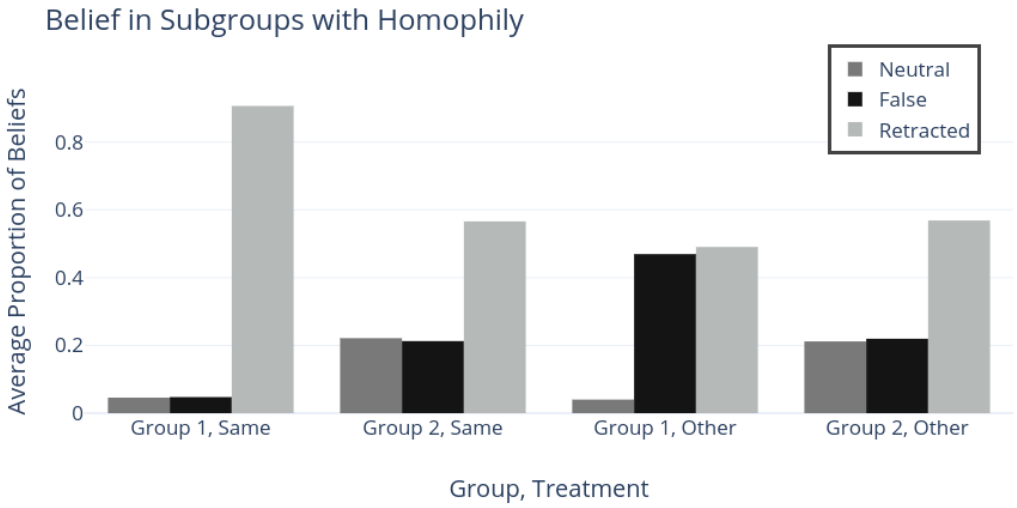


Figure 12: Average length of belief for two subgroups in a homophilic network. Results are for group 1 and group 2, for cases where the retraction is introduced in the same subgroup (1) and the other subgroup (2). $p_{in} = 0.4, p_{out} = 0.004$, $N = 100$, Delay=400, timed novelty stops after 200 rounds

To summarise, 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. They show how network structure can impact the success of retraction, and especially why homophily might lead to issues with retraction.

4. Conclusion

Simplified models like the ones we present here must always be treated carefully when applied to real-world cases like retraction in epistemic communities. 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, and how they are useful to philosophy of science.

One of these gaps is that we do not model prominent, or central communicating agents such as a journal, or academic search engine, that will influence or 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 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 in the real world. 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—i.e., agents may think it likely, but not sure, that some belief is true or false. 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 in our populations, 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. They do not exhibit anchoring or recency bias. Of course, in many real cases, these effects will be at play. For this reason, as suggested earlier, these models should be applied only to cases that best fit their assumptions. Further work might look at whether the results presented here hold up under more complex representations of belief states, and the addition of reasoning biases.

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 that scientists often continue to cite and share their own retracted findings. Madlock-Brown and Eichmann (2015) use citation data to show that this is a common practice and that scientists who do it boost their post-retraction citation count. Why would authors do this given norms that forbid the sharing of false information in science? As the sociologist of science Robert Merton convincingly argued, scientists are often motivated by *credit* (Merton, 1973)—attention and good reputation from those in the community and all the positive benefits that follow, such as prestigious invitations, higher pay, grant funding, awards, etc. This may help explain why individual scientists continue to cite and promote retracted work.

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).[17] 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 more thorough 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 who have political or economic motivations to shape public belief. Besides credit-motivated scientists, we might also include agents with economic or political motives who seek to shape the landscape of belief. Philosophers of science have already used network epistemology models to investigate situations like

[17]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).

this, finding that agents who intervene in networks to promote false beliefs can be very effective at doing so (Holman and Bruner, 2015, 2017; Weatherall et al., 2018). Contagion models might provide a constructive addition to the growing literature.

What are appropriate takeaways given the limitations of these models? There are several. First, they are useful to hypothesis generation, and the direction of 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. What solutions do they suggest? We cannot actually 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 of time than they usually would. However, we might be able to institute changes with respect to 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 retracted information actively and indefinitely. For this to work, real organisations should be more direct 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 where appropriate. This would, in effect, be a policy designed to promote continued and

widespread sharing of the retraction.

Our results also suggest that it is important for retractions to be spread by the same sources that originated a false belief. This means that if a refutation to some result is published in another field or subfield, it might not be effective. The journal that originally published the false result should consider publicising 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.

Acknowledgements

[Removed for review]

References

Adar, Eytan and Lada A. Adamic (2005). "Tracking Information Epidemics in Blogspace." *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. IEEE Computer Society, 207–214.

Bala, Venkatesh and Sanjeev Goyal (1998). "Learning from neighbours." *The review of economic studies*, *65*(3), 595–621.

Baldwin, J. Mark (1894). "Imitation: A chapter in the natural history of consciousness." *Mind*, *3*(9), 26–55.

Begley, C Glenn and Lee M Ellis (2012). "Drug development: Raise standards for preclinical cancer research." *Nature*, *483*(7391), 531.

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.

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.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. (2016). "Evaluating replicability of laboratory experiments in economics." *Science*, *351*(6280), 1433–1436.

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

Collaboration, Open Science et al. (2015). "Estimating the reproducibility of psychological science." *Science*, *349*(6251), aac4716.

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

De Tarde, Gabriel (1903). *The laws of imitation*. H. Holt.

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.

Frey, Daniel and Dunja Šešelja (2018). "Robustness and Idealizations in Agent-Based Models of Scientific Interaction." *The British Journal for the Philosophy of Science*, *axy039*.

Garetto, Michele, Weibo Gong, and Don Towsley (2003). "Modeling Malware Spreading Dynamics." *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*. IEEE, 1869–1879.

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.

Hudson, John (2012). *Yes Roger, Fox News Has Retracted False Stories*.

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

Ioannidis, John PA (2005). "Contradicted and initially stronger effects in highly cited clinical research." *Jama*, *294*(2), 218–228.

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

Kim, Louis, Mark Abramson, Kimon Drakopoulos, Stephan Kolitz, and Asu Ozdaglar (2014). "Estimating Social Network Structure and Propagation Dynamics for an Infectious Disease." *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 85–93.

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

Le Bon, Gustave (1896). *Psychologie des foules [Psychology of the Crowd]*. F. Alcan.

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.

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.

Mayo-Wilson, Conor, Kevin JS Zollman, and David Danks (2011). "The independence thesis: When individual and social epistemology diverge." *Philosophy of Science*, *78*(4), 653–677.

McWilliams, James (2013). *Journalism is Never Perfect: The Politics of Story Corrections and Retractions*.

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

Moore, R. A., S. Derry, and H. J. McQuay (2010). "Fraud or Flawed: Adverse Impact of Fabricated or Poor Quality Research." *Anaesthesia*, *65*(4), 327–330.

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.

Prasad, Vinay, Andrae Vandross, Caitlin Toomey, Michael Cheung, Jason Rho, Steven Quinn, Satish Jacob Chacko, Durga Borkar, Victor Gall, Senthil Selvaraj, et al. (2013). "A decade of reversal: an analysis of 146 contradicted medical practices." *Mayo Clinic Proceedings*. Elsevier, 790–798.

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

Rosenstock, Sarita, Justin Bruner, and Cailin O'Connor (2017). "In Epistemic Networks, Is Less Really More?." *Philosophy of Science*, *84*(2), 234–252.

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.

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. "Even retracted papers endure." *Science* 312, 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.

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*.

White, Paul, Henrik Kehlet, and Spencer Liu (2009). "Perioperative Analgesia: What Do We Still Know?." *Anesthesia & Analgesia*, *108*(5), 1364–1367.

Young, H Peyton (2009). "Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning." *American economic review*, *99*(5), 1899–1924.

Zollman, Kevin JS (2007). "The communication structure of epistemic communities." *Philosophy of science*, *74*(5), 574–587.

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

Appendix A. Proofs

Proof of Proposition 1. *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. Further, by the set-up of the model, at least one individual holds the retraction. By assumption, there is always positive probability that any two individuals are paired for interaction. When this happens, the state configuration will change because the individual with the false belief will adopt the retraction. Therefore, the state is not stable. □