# FROM EVIDENCE TO OPINIONS:

# EXPLORING VARIANCE IN SCIENTIFIC COMMUNITIES

A Dissertation

by

SEAN ROBERT CONTE

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

| | |
|---|---|
| Chair of Committee, | Kenneth Easwaran |
| Committee Members, | Martin Peterson |
| | Nathan Howard |
| | Yoonsuck Choe |
| Head of Department, | Theodore George |

August 2023

Major Subject: Philosophy

ABSTRACT

Much of science depends on random sampling, and random sampling always involves *variance* — samples of the same population can have different results. In this dissertation, I explore how variance between samples manifests as variance between scientists' opinions as well as unpredictability of the community as a whole, depending on how the community handles the samples. I show choices between network structures, strategies for sharing information, and strategies for trusting information impose trade-offs concerning the goals that a community might have. Finally, I consider how the variance of studies might be distorted by two forms of inherent reliability of scientists, imprecision and bias, and evaluate solutions to the resulting challenges.

To conduct this exploration, I design computer models to simulate scientific research. Scientists update their opinions using data that they produce individually and share amongst each other. I model core aspects of science and bracket away further complications, which obfuscate underlying dynamics. By isolating particular features of scientific communities in this way, I am able to identify novel dynamics inherent to science on a fundamental level.

In particular, I show how certain ways of diversifying opinions of a community can be *useful* by enhancing the predictability of that community. I follow this by uncovering how a community can achieve predictability *without* diversifying opinions. This sets up an investigation into such methods as potential solutions to severe challenges of scientific communities that are caused by data manipulation and insidious forms of information sharing.

This project provides a foundation for how to approach distributing statistically generated evidence (i.e. random samples) on networks. One might think that interesting questions concerning information sharing only emerge for complicated environments and that sharing scientific information alone is largely trivial. I show that this intuition is false. I show there are fundamental dynamics concerning how statistically generated data is shared, and exploring these is vital for understanding the more complex situations.

DEDICATION

To scientists, or at least, some of them...

# ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Kenny Easwaran, who taught me to be analytic without being reductive and showed me that being curious is more important than being certain. I would also like to thank my undergraduate advisor, Dr. Dugald Owen, who first showed me logic and philosophy and gave me the tools for thinking clearly. Thank you also to my committee members, Dr. Martin Peterson, Dr. Nathan Howard, and Dr. Yoonsuck Choe, without whose teachings this project would never be possible.

I owe a great deal to my friends through the years, Jared Oliphint, Alexandar Crist, Jyothis James, David Anderson, Michael Portal, Haley Burke, Victoria Green, Anton Classen, Rachel Cicoria, Brandon Wadlington, Denise Meda-Lambru, Kristian Cantens, Luca DiVincenzo, Ethan Pak-Him Lai, Abouzar Moradian, Firooz Jifari, Nick Charles, Brady DeHoust, Gedalyahu Wittow, Alex Haitos, Chris Black, Ryan Manley, Matt Wester, Wendy Bustamante, Dong An, and James Reed. The conversations we had played an indispensable role in how I learned to think about diverse epistemologies.

Thank you also to my students, who are too numerous to name. Teaching and learning with you all was one of the most enjoyable aspects of this experience.

This project would not have been possible without my family. To Mom and Dad, thank you for making this possible at every step of the way. The world would have few problems if all parents were as loving and supportive. Thank you to my siblings, Kath, David, Brandon, Chris, and Mark. Because of you all, my analytic disposition began earlier than most. I am also indebted to my growing family, Libby, Brynn, Jeri, David, Cheryl, Mike, Sabrina, Jerald, and Shay who have provided me support I didn't even know I needed.

Most of all, thank you to my wife, Helen, whose strength, compassion, and love has kept me afloat throughout this process. Everything I have accomplished and will accomplish is made possible by you.

# CONTRIBUTORS AND FUNDING SOURCES

## Contributors

This work was supported by a dissertation committee consisting of Professors Kenny Easwaran (advisor), Martin Peterson, Nathan Howard, and Yoonsuck Choe.

All work conducted for the thesis dissertation was completed by the student independently.

## Funding Sources

No outside source of funding was provided.

# NOMENCLATURE

The following nomenclature is relevant to the main model I develop:

| | |
|---|---|
| $N$ | Number of scientists in the community |
| Sample/Study | A set of data points taken by a scientist to determine a particular frequency |
| $0.5 + \epsilon$ | The true frequency in question |
| $n$ | The size of each scientist's sample |
| $k$ | The number of successes a scientist observes |
| $\frac{k}{n}$ | The observed frequency of a particular sample |
| Network | Indicates with who agents are able to share information |
| Cyclic Network | Each agent has two connections, no two agents have a third in common (for $N > 3$) |
| Complete Network | Each agent is connected to every other agent |
| Wheel Network | All agents except one form a cyclic network; the remaining agent is connected to everyone |
| Star Network | One agent is connected to all the rest, and there are no other connections |
| Star Agents | The agents on a network connected to everyone |
| Perimeter Agents | The non-star agents of a network (thought of as situated in a ring around the star agents) |
| Cycles | The number of nearest neighbors that perimeter agents are connected to |
| Imprecision | Distortion which might skew a samples frequency up or down, *irrespective of* the researcher's opinion |
| $I$ | The maximum level of imprecision in the community |
| $i_j$ | The level of imprecision of agent $j$ |

| | |
|---|---|
| Bias | Distortion which might skew a samples frequency up or down, *depending on* the researcher's opinion |
| $B$ | The maximum level of bias in the community |
| $b_j$ | The level of bias of agent $j$ |
| Disclosure Strategies | The rule by which agents share the information they produce |
| Full-Disclosure | Share all of the studies that you produce |
| Selective-Disclosure | Share the studies that you produce which you believe reflect the correct opinion |
| Non-Disclosure | Share none of the studies |
| Trust | The amount that an agent's testimony is discounted |
| $w_{jl}$ | The amount that agent $j$ trusts agent $l$ |
| Trust Strategies | The way that an agent's trust is adjusted |
| Credulism | Agents trust each other fully |
| Skepticism | Agents do not trust each other at all |
| Homophily | Agents trust each other in direct proportion with the extent that they agree |
| Heterophily | Agents trust each other in direct proportion with the extent that they disagree |
| Imprecision-Avoidance | Agents trust each other less the more imprecise they are |
| Bias-Avoidance | Agents trust each other less the more biased they are |
| Round | A single execution of the model in which agents research, share and update on their information |
| Run | Consecutive rounds of the model in which agents acquire data over time |
| Simulation | Iterating a set of runs with the same parameter settings numerous times |
| Estimate | The total amount of $k$ over $n$ that an agent has seen up to a given round |

| | |
|---|---|
| Credence | The confidence level of an agent that the frequency in question is above 0.5 represented on a $[0, 1]$ scale |
| Logodd | The confidence level of an agent that the frequency in question is above 0.5 represented on a $(-\infty, \infty)$ scale |
| Aggregate Opinion | The average opinion of every member of the community; can be defined in terms of estimate, credence, or logodd (when analysis does not apply to all three, they will be specified) |
| Speed | The rate at which the aggregate opinion increases over time |
| Predictability | The variance between the aggregate opinions on different runs of a simulation |
| Diversity of Opinions | The variance between individual opinions within a single run |
| Diversity of Evidence Produced | The variance of all the evidence that is generated in the community (regardless of who produced it or who will update on it) |
| Diversity of Evidence Held | The correlation between pools of evidence that agents in the community hold |
| Diversity of Research Paths | That agents study different options (does not apply to my models) |
| Preferential Research | The stipulation that agents research is contingent on their opinion (does not apply to my models) |
| Siloing | When agents generate evidence from identical distributions but temporarily diverge in their opinions due to independent pools of evidence |
| Extremetization | When pools of evidence are correlated, causing agents to pull each other further from the average than they otherwise would go |
| Feedback-loops | A increase (decrease) in opinion *always* makes a peer more likely that one's peer will increase (decrease) and *vice versa* |
| Echo-chambers | A set of agents engaged in a feedback-loop |

| | |
|---|---|
| Group-think | A community consisting of one echo-chamber that unanimously reassures itself of whatever answer it as arbitrarily selected |
| Entrenchment | A community fracture into echo-chambers that perpetually move away from each other |
| Intermediate Agents | Agents connected to, but not a part of, disagreeing echo-chambers |

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

# 1. AN INTRODUCTION TO EXPLORING VARIANCE IN SCIENTIFIC COMMUNITIES

## 1.1 Introduction

Suppose you want to find out the frequency of something. For a few examples, suppose you want to know how successful a medical treatment is, how common a particular substance is, or the correlation between social-economic conditions and outcomes. Determining the frequency of any natural quantity requires taking samples. Doctors administer a treatment to a group of patients and observe how many of them are cured. Geologists take deposit samples from their research site and test for the presence of certain elements. Social scientists might poll a random subset of a population. These are just a few of the countless examples where science depends in some way or another on taking samples, setting aside how vastly different the fields may be otherwise. Researchers in each of these fields use their samples to inform their opinions concerning their frequency (success rate, prevalence, etc.) of interest.

Sampling inherently involves *variance* — any given sample likely has a success rate above or below the overall or actual success rate. Even when groups of patients are given the same treatment, it might be that more are cured in one than the other. That is why researchers take multiple samples, so that over time the aggregate observed frequency reflects the true frequency. So far, this aspect of variance has been well appreciated by scientists. However there is an important effect of variance between samples that has so far been largely ignored: that the *variance of evidence* leads to a *variance of opinions*. When two researchers produce and update on samples that have different frequencies, this leads to a difference in their opinions. Therefore, this is a case in which evidence itself dictates that researchers disagree. One might think that researchers sharing evidence is the obvious solution to this disagreement, but in this dissertation, I show that that depends on one's goals. In this dissertation, I explore the effects of variance in scientific research on scientific opinions.

I consider different aspects of the community and how these interact with the variance of studies to affect multiple metrics of performance including the diversity between their opinions and the predictability that they will have a certain opinion given a certain amount of evidence. For a few examples, I consider the amount of connections between researchers, how evenly spread out those connections are, the trusts strategies between researchers, the disclosure strategies of researchers, as well as the potential for some to be less reliable than others. In each of these cases, I provide insights into how the issues caused by variance of scientific evidence are far from trivial. I show the different decisions that communities have for dealing with variance impose trade-offs between epistemic desiderata.

In addition to providing a foundation for an exploration of data variance in communities, this dissertation provides a richer understanding of each of the following concepts: epistemic diversity/polarization, the predictability of a community, the connectivity of a network, the centrality of a network, trust strategies, disclosure strategies, the imprecision of scientists, the bias of scientists, siloing, extremetization, echo-chambers, entrenchment, group-think, and the difference between group and individual epistemic desiderata.

This exploration is conducted with the use of computer models which involve making dischargeable assumptions to allow for isolating the dynamics of the systems in question. Some of these assumptions are made to remove complications, e.g. that agents do not make mathematical errors. While these assumptions negate certainly real aspects of many real communities, such aspects obfuscate more fundamental dynamics. These assumptions are tantamount to physicists assuming a frictionless surface. Doing so lets us focus on dynamics that might be complicated or obfuscated in more complex environments, but which hold nonetheless. Other assumptions, like Bayesianism (which concerns one way in which agents update on evidence), are made to facilitate looking at different metrics (such as credences), but they play no role in the explanation of the findings. The same dynamics can be seen with completely non-Bayesian metrics.

I analyze the community with a variety of metrics, because doing so provides a much

richer understanding of the dynamics themselves. None of the metrics should be taken to exhaustively capture the performance of the community — even with regard to a particular dynamic. Each metric provides another lens by which the dynamics of the community can be viewed. No particular lens tells the whole story, and likewise particular dynamics should not be *identified* as performance on some metric. The dynamics discovered can be observed by many metrics, so the dynamics are never *defined* in terms meeting some threshold on a particular metric.

Similarly, I do not assume what the right epistemic desiderata are. The dynamics I uncover are relevant regardless of what turns out to be important. This project does not defend normative claims. Similar to economic models, it makes *descriptive findings that inform* normative questions.

In what follows I discuss the relevant literature followed by a discussion of my base model. This is followed by a brief overview of the chapter findings.

### 1.2 Background

There are a number of projects related to mine. This section explains how mine is situated amongst them. There are numerous topics with even more numerous connections between them. I have broken them up into four headings: divisions of epistemic labor, sharing information, echo-chambers, and disagreement & trust.

The origins of this project can be traced to philosophers of science like Kuhn who focused heavily on science as an inherently social enterprise. Kuhn (1970) focuses on how societies transition through periods of normal versus revolutionary science. These periods are marked by whether or not communities belong to particular paradigms, i.e. when they share values as well as beliefs and theories. Revolutionary science is the product of significant dissent among scientists.[1] While I am not specifically concerned with scientific revolutions here, what I say does have significant impact on how scientists might diverge in opinions concerning particular issues. In future research I plan to build on this to understand elements of paradigm shifts.

[1]More recently Strevens, 2020 also focuses on the social elements inherent to the development of science.

Kuhn, 1970

### 1.2.1 Divisions of Cognitive Labor

More recently, Kitcher (1990) argues for a *division of cognitive labor* in the sciences. Kitcher highlights the importance of specialization in multiple ways like scientists having expertise in particular fields, scientists in a particular field testing different theories, as well as different scientists in the same lab having different jobs. This has since lead to multiple instances of philosophers using *computer models* to explore divisions of labor in different ways. So far, these largely focus on how to ensure that sufficient evidence is produced to guarantee that the community has found the optimal option or the most accurate opinions. Two general methods are worth looking at before seeing how my project relates.

One way in which philosophers model divisions of labor is in regards to how willing someone should be to try untested (high risk, high reward) options versus those that they already know but are potentially not optimal (low risk, low reward). This is known as the *explore-exploit* dilemma: should an agent exploit their known alternative or explore new ones. Weisberg and Muldoon, 2009 describe two types of scientists: *mavericks*, who are more inclined to test more unknown theories, and *followers* who are more inclined to test theories that have already been explored. The authors show that the communities do best with regard to finding the optimal theory when there is a mix of both types of agents.

A different way of promoting a diversity of research paths concerns diversifying the agents' opinions themselves (O'Connor and Gabriel, 2022; Wu, 2022; K. J. S. Zollman, 2007, 2010[2]). These projects focus on how to generate a division of research by inducing a division of opinions/research interests. The most famous result among them is called the *Zollman Effect*: that lower connectivity (and communication) can lead to more diverse opinions concerning the options open to scientists. As a result, the scientists are more likely to explore all of their alternatives and have a greater chance of finding the right one.[3]

[2]This is well summarized in O'Connor and Wu, 2021

[3]This is also related to a debate concerning whether or not there are any cases where diversity trumps ability (Grim et al., 2019; Hong and Page, 2004; Singer, 2019; Thompson, 2014). I do not contribute to that

In contrast to the above to methods, I explore the effects of a distribution of evidence. The division of epistemic labor that I am interested in is how to divide up who is responsible for updating on what evidence. I am interested in the flow of information, not whether or not it is generated. The above projects show countless important phenomena, but they all involve too many complications for us to understand variance of data on networks at its core. On my models, every agent researches every option every round. Therefore there is a constant amount of information produced. The observed effects only concern how that information is handled. So unlike the above models, where the notion of epistemic diversity is tied to differences in research predilections or opinions, in my models the epistemic states of the agents do not affect whether or not they research. I use multiple ways to capture the opinions of the agents, but none of them affect if data is generated. While the above projects investigate the value of diverse opinions as they might lead to diverse paths of research, I am interested in diverse opinions as a reflection of diverse evidence. As I discuss later, I generalize the Zollman Effect by showing that decreasing communication enhances predictability even when all agents are committed to research. One upshot of all this is a clean distinction between the diversity of evidence, diversity of opinions, and diversity of research.

### 1.2.2 Sharing Information

The Zollman Effect is related to what is known as the *communist norm* in science, which is that scientists ought to share all of their information (scientists should not withold results from others). Works like Heesen, 2017a and Strevens, 2020 discuss models where they evaluate if value-maximizing agents will abide by the norm. I am also concerned with how/when/if evidence should be shared, but I investigate this in a different way. I am concerned with determining if the communist norm (in its many manifestations) is useful for

debate directly, because doing so would require analyzing their specific models and inferences made, and that is far afield from what I am doing here. However, my results do our contribute to our understanding of epistemic diversity more generally, and I make distinctions which are likely very helpful for that debate (especially the difference between diversity of evidence and of opinions).

a community generally, and based on those findings discuss whether the individual agents have a reason to care.

There are some projects concerned with community level effects of sharing. For instance, Heesen, 2017b, 2018; Heesen and Bright, 2021; Soysal, 2019; K. Zollman, 2009 are concerned with the strategies that publishers take for what and how information is disseminated. Similarly, Weatherall et al., 2020 concerns propagandists who share scientific data. As such, each of these projects are specifically concerned with institutions whose sole function is that of sharing information. This typically involves a number of assumptions that are unique to such cases (like having the ability to see all of what is produced in the society). Conversely, my project concerns how the agents themselves should share their information. Not all communities include media institutions, but all communities are faced with deciding how its agents ought to share their information. I start at this fundamental level and work my way up to more complicated sharing situations.[4]

The way in which agents share information is also related to the literature on *information cascades* and *jury theorems*. Hung and Plott, 2001 famously shows how information cascades can form in communities of rational agents. That is, depending on the structure of how information is reported, agents might be convinced in the wrong answer even though in isolation most agents' evidence indicates the correct answer. Alternatively, List and Goodin, 2001 shows that (when reporting is done right) agents need only be minimally reliable ($>$ 50%) to ensure they find the truth. Hence we can be confident in group reliability of juries even with relatively weak reliability of members. I do not directly consider the same models or structures of disclosure, but my results contribute to this discussion by showing other ways in which the flow of information can lead to more or less predictability. I show that the wrong ways of sharing can lead to *feedback loops* which can manifest as *group-think* in a very similar way to information cascades, while the right forms can lead to guaranteed success

[4]For an example of building to more complicated situations, I consider agents with more influence than others based on network structure. Such agents have access to (and also share) more information than their peers.

indicative of a reliable jury. As such, my project offers a way to consider the relationships between information cascades and jury theorems in a new environment.

### 1.2.3 Echo-Chambers

Feedback loops play an integral role in how I characterize *echo-chambers*. Echo-chambers can persist even when its members are exposed to alternative information. What is important is that agents are engaged in a feedback loop where they perpetually reinforce each other's opinions (inspite of any other information they hear). Other works, like Nguyen, 2020, characterize echo-chambers as "epistemic bubbles" where agents are ignorant of information outside their bubble. I refer to this as *siloing*. Both feedback loops and siloing play an integral role in echo-chambers, but only the former are necessary for their existence. Feedback loops are at the core of what an echo-chamber is, while siloing determines how that echo-chamber will manifest, i.e. how fractured the echo-chambers of a community will be (more on this shortly).

Aside from sharing information, echo-chambers can also be caused by *bias* among agents. There are multiple projects that look at different forms of bias (Holman and Bruner, 2015; Holman and Geislar, 2018; Lord et al., 1979; O'Connor and Gabriel, 2022). I am particularly concerned with a form of bias discussed in Bright, 2021 in which agents distort their evidence to support their view. This can take place in the form of researchers overtly changing their results, or it might be due to subconscious errors in how observations are encoded. Either way, bias of this sort is all too simple yet prevalent to not explore thoroughly. My project contributes to the existing literature on bias by providing an in depth investigation of this (likely) highly prevalent, highly effective, yet highly ignored version. In addition, it provides a framework within which to recreate the other forms of bias (e.g. distortions on how one updates on their evidence) so that their effects can be compared.

However echo-chambers, and likewise bias, do not always cause group-think, but instead (depending on the level of siloing) sometimes lead to *polarization*, another concept which has received considerable attention recently (Bramson et al., 2016; Haghtalab et al., 2020;

O'Connor and Weatherall, 2018; Singer et al., 2019; Weatherall and O'Connor, 2020). In contrast to the forms of diversity above (to which philosophers attribute some instrumental value), polarization is a diversity of opinions that is considered undesirable. Many of these projects are about unearthing what mechanisms cause polarization. I contribute to this by identifying a particular form of polarization, *entrenchment*, where communities fracture and perpetually dig in their heels, and I find two independently sufficient causes for it.

### 1.2.4 Disagreement & Trust

A related discussion is how agents should react in light of contrasting opinions or evidence (Bicchieri, 2005; Fricker, 1994; Lackey, 2007, 2008). For example, Easwaran et al., 2016 considers different ways in which agents might adjust their credences in light of peers having different credences. Contrarily, I do not consider agents that react to each other's *credences*, but only each other's *data*. This project is about information flow, and credences are just one of many lenses with which to capture the effects. So I do not contribute directly to the discussion for how agents should adjust their credences based on those of others. But I do explain how disagreement can be navigated by sharing information, and I set the stage for other projects which do include changing credences based on other credences.

A similar discussion to that of peer disagreement concerns how agents should trust each other. Some philosophers apply computer models in evaluating what they call "norms of trust" though I prefer the term *trust strategy*. Mayo-Wilson, 2014 as well as K. J. S. Zollman, 2015 investigate different ways in which agents might put more or less stock into their peers' respective testimonies. My project contributes directly to these by investigating even more trust strategies as well as metrics for evaluating them. I consider important new metrics and show how a strategy that has so far not been considered performs best with regard to them. Further, I investigate these trust strategies in the context of epistemic challenges like imprecise or biased scientists. I show how trust strategies are impacted differently based on the presence of various challenges.

One important strategy for trusting others is called *homophily*, and it has received con-

siderable attention recently (Fazelpour and Steel, 2022; Golub and Jackson, 2012; Mohseni and Williams, 2021; K. J. Zollman, 2012). This is the tendency for agents to associate (or trust) those that are similar to themselves. Homophily has been studied for both its positive and negative impacts, but what has gone largely ignored is its counterpart: *heterophily*. Heterophily is the tendency to associate with those *dissimilar* to oneself. I expand the discussion of homophily by considering this important alternative. In doing so I not only provide alternative (and more useful) options for agents, but I also add to the understanding of homophily itself.

## 1.3 Model

### 1.3.1 Setup

My models capture scientific research concerned with generating random samples. In this section, I describe the base version of a model that I use throughout the chapters. The model concerns a network of agents that share statistically generated evidence to update their opinions concerning a particular problem. The value of this model is that it applies to the broad range of scientific communities that rely on random sampling, and it depends on minimal epistemic assumptions. I will first explain the basics of the model before the metrics used to explore the model.

The problem in question can be thought of as whether or not a particular frequency in nature is above or below a given threshold. One can think of this as doctors determining if a medical treatment's success rate is higher than a known standard, but there are many interpretations of the model. All that is important is that the agents use statistically generated evidence to update their opinions. The community consists of $N$ many agents that each produce $n$ new data points every round. The observed frequency of a study is $\frac{k}{n}$ where $k$ denotes the number of successes for the study. The $\frac{k}{n}$ of any given study is likely above or below the true frequency, but overtime the aggregate frequency of the studies reflects it. Agents research every round. This ensures the amount of evidence that is produced each

round is constant, which implies that the dynamics that unfold are due purely to the flow of information and not how much is produced.

Every round $N \times n$ many data points are produced in the community, but the amount of those data points that any given member of the community observes is dependent on the network structure. Agents only observe the studies of those they are connected to on the network. One of the topics I investigate is the effects of different parameters of the network on the performance of the community.

Each agent has an opinion which can be captured in different ways. The most basic is their *estimate*, i.e. the total number of successes divided by the total number of trials. The results in this dissertation can be understood with regard to this measure alone, but it turns out to be helpful to include others. For instance the *credence* of an agent is a number between 0 and 1 and indicates their level of confidence in the claim that the new treatment is actually better. Likewise their *logodd* captures the same quantity but on a logarithmic scale. Both are determined in a way consistent with Bayes' rule. However it is important to note that Bayes rule does not play a role in any of the explanations. These extra lenses merely provide more ways to capture the dynamics in question.

### 1.3.2 Metrics

I use multiple metrics which each depend on the opinions of the individuals. Versions of each metric can be calculated in terms of either estimates, credences or logodds, so I will define them in terms of opinions themselves. The *aggregate opinion* of the community is the average opinion of all of the community members. I guage the aggregate opinion in two ways: First, I am concerned with how quickly the aggregate opinion changes (*speed*). That is, given a certain amount of evidence, how confident (and likewise accurate) can the members of a community expect to be on average. Second, I am concerned with the how predictable the aggregate opinion is (*predictability*). That is, how similarly do different runs of the simulation perform. In addition to the aggregate of the agents, I am also interested in the variance between the individual agents' opinions (*diversity*).

The point of these metrics is not to assume anything about what the right epistemic desiderata are, but instead these projects reveal mechanisms that are relevant to whatever turns out to be important epistemically. The insights of value are never exhaustively captured by any metrics, and likewise no dynamics should ever be identified with the performance on particular metrics. In the philosophy of mind, there's a slogan: "Thought is first, language is second." Something similar is true here: *Dynamics first, metrics second.*

### 1.4 Results

Chapter 1 shows that decreasing the connectivity of a community can increase its predictability. More specifically, it details how increasing communication limits the possibility of variance between opinions. As a result, the more highly connected a community is, the more it will be influenced by outlier samples. Meanwhile, more sparsely connected communities are more insulated against outliers. This demonstrates a way in which diversifying evidence can increase the predictability of a community. In doing so, it generalizes the Zollman Effect discussed above by showing that the effects of diversity are not limited to situations where agents might abandon research. Diversity's effects on the predictability of opinions concern the flow of information alone. When research is contingent on these opinions (like in Zollman's and others' models), this unpredictability is manifested as a greater potential to abandon research.

Chapter 2 explores what other sorts of diversity imbuing changes lead to predictability. While Chapter 1 finds one change (decreasing connectivity) which, by causing diversity, increases predictability, not everything that causes diversity makes a community more predictable. I falsify two hypotheses concerning what guarantees that a diversity-imbuing change leads to a more predictable community. I defend a third hypothesis which identifies the appreciation of the variance of evidence as the fundamental mechanic in what makes diversifying opinions lead to a more predictable community (when it does). That is, increased predictability is guaranteed when a change makes it so that a community's diversity (variance) of opinions more closely reflects the variance of the data that they generate

individually.

Chapter 3 considers how a community might use trust strategies to increase their predictability *without* increasing their diversity. While Chapter 2 uncovers when diversity increases predictability, Chapter 3 shows that there are other ways to do so. Hence increased diversity is not a necessary condition for increasing predictability. I show how a community can do so with heterophily. Using that strategy, agents are able to avoid siloing (increased variance due to independence between results of disagreeing agents) as well as extremetization (the correlation between results of agreeing agents).

Chapter 4 concerns how bias (in this case, manipulating data) can lead to echo-chambers for a community. Based on the level of siloing, these echo-chambers either manifest as group-think or entrenchment. In addition, I show how various trust strategies mitigate and exacerbate the echo-chambers. In particular, I show how homophily intensifies them, while heterophily shows promise for a possible solution. This chapter simultaneously explores an under-explored form of bias, provides a new characterization of echo-chambers, group-think, and entrenchment, and finally, shows how heterophily can be far more useful than homophily in important cases.

Chapter 5 investigates the concept of selective-disclosure, which one might think would lead to the best epistemic ends. The strategy involves only sharing studies that one thinks reflects the correct opinion, and not sharing any that one deems to be flukes (ones that would cause a less accurate credence). I show that this strategy can have the same impacts as manipulating information, and as a result can lead to the same echo-chambers discussed in Chapter 4. In addition to showing that the ways in which information is shared is far from trivial, I show that heterophily has an even better time mitigating echo-chambers in these cases.

The upshot of this research is a much deeper understanding of both 'good' and 'bad' forms of diversity. I show how dynamics which contribute to the propagation of diversity can make

a community more predictable, and I show communities can also use heterophily to curtail diversity in a way that doesn't lead to unpredictabiltiy. In doing so, I provide a foundation for understanding how information flows on epistemic networks.

# 2. GENERALIZING THE ZOLLMAN EFFECT: REVEALING THE UNEXPLORED UTILITY OF OPINION DIVERSITY

## 2.1 Overview

This chapter reveals a previously ignored utility of diversifying opinions. It shows that limiting connectivity between scientists that share statistically generated evidence increases the predictability of the community as a whole. This dynamic occurs purely based on the flow of information itself, and the fact that increasing communication increases the amount of information *consumed* without increasing the amount *produced*. Many projects, like K. J. S. Zollman, 2007 which also considers amounts of connectivity, argue that promoting a diversity of opinions can be beneficial because it ensures that all possible research paths are sufficiently explored. But I show that there is a usefulness of diversity that is independent of its impact on research paths chosen. I generalize Zollman's results and uncover a more fundamental dynamic. Namely, increasing diversity by decreasing connections causes a community to better reflect the diversity of the evidence itself.

## 2.2 Introduction

I show that decreasing communication of scientific data causes the community in question to become more predictable. When scientists in a particular field are researching the same issue, their opinions on the issue depend heavily on the amount of communication between them. It is tempting to think that sharing as much evidence as possible (thereby curtailing the diversity of opinions) is the best way to achieve epistemic desiderata, as it allows the scientists more (legitimate) information on which to base their opinions. However there are many epistemic desiderata. With the use of two computer models, I show that something like this intuitive story is true for some of them, but not all. I show that the amount of communication between researchers imposes a trade-off between how quickly the community aggregate opinion becomes more accurate (*speed*), the variance of the aggregate opinion between simulations (*predictability*), and the spread between their individual opinions within a simulation (*diversity*).

Other projects have made similar findings, but with a crucial difference. Most famously, K. J. S. Zollman, 2007, 2010 show that when scientists have too much communication, they can be uniformly misled by studies that indicate the suboptimal theories are better. This can cause them to abandon researching the optimal theory and likewise never realize it is actually best. This result has been come to be known as the *Zollman Effect*, and it has had significant influence.[1] It is a paradigmatic example of projects in the literature which attribute value to diversity of opinion. Zollman, like many other projects (O'Connor and Gabriel, 2022; Wu, 2022), attributes value to diversifying opinions when doing so also diversifies research paths. These projects differ by the ways that they promote diversity of opinions (e.g. level of connectivity, stubbornness, cultural divisions), but the goal is to find such implementations worthwhile (or at least beneficial in some cases) due to their ability to promote the right kind of diversity. This chapter generalizes the Zollman Effect by showing

[1] K. J. S. Zollman, 2010, which focuses primarily on this result, has been cited by nearly 200 other projects so far.

how a diversity of opinions can make a community more predictable even without impacting what is researched. I uncover a dynamic fundamental to sharing statistically generated data, by showing that in communities where such data is shared: decreasing communication increases the predictability of the community.

This result has two important implications: First, it provides a better understanding of connectivity itself. Increasing the number of connections increases the amount of data consumed without increasing the amount produced, and doing so is not necessarily the right strategy for every community. Second, it establishes a so far ignored aspect of the value of epistemic diversity. While other philosophers attribute value to diverse opinions as a mechanism for yielding diverse research plans, I show diversity of opinions can be useful even when it does not affect what is researched. That is, I show diversity of opinions increases the predictability of a community without affecting the agents' research itself. This chapter focuses on how the amount of connectivity can influence diversity, and the ways which this might be good or bad. In the next chapter I consider other ways of causing diversity, and pin down exactly what differentiates useful causes of diversity from purely detrimental ones. This project helps expose a mechanic which is defended in the next chapter as a sufficient condition for when diversifying opinions is useful: when it reflects the diversity of research produced. Here I reject the necessary condition that in order for diversity of opinion to be valuable, it must be in promotion of a diversity of research paths.

The chapter proceeds by first recapitulating Zollman's results, i.e. the Zollman Effect as it has come to be known, before laying out the concerns that one might have with the Zollman Effect. Next, I show that a generalized version Zollman Effect holds using a variety of metrics and with two models. In doing so I alleviate the concerns that one might have about the Zollman Effect and also provide a deeper explanation of it.

## 2.3 Background

Zollman makes his finding using a model that is adapted from one originally designed by economists Bala & Goyal (1998), which itself has a result that resembles (what is later

known as) the Zollman Effect. The finer details of their model are not important, but it is helpful to have an idea of how the relevant features lead to something which resembles what Zollman finds later.

### 2.3.1 Bala & Goyal

Bala & Goyal's model consists of an infinite number of agents that are attempting to find the optimal strategy for a particular problem. They individually employ different strategies, and over successive rounds, learn about the quality of their strategies based on their own performance and that of their neighbors on a network. Research is preferential: the agents decide which strategy to try based on their own performance and that of their neighbors given each agent's respective strategy. The authors compare two networks with regard to the agents' ability to converge on the optimal strategy for a given situation. On the first network, the agents have on average three connections apiece. On the second network, each agent has two connections.

The authors show that the more connected community is quicker at converging on an answer, but only on the less connected community are the members guaranteed to find the optimal strategy. This is because the greater connectivity makes it possible that the agents too quickly converge on a suboptimal strategy, leaving no one to test the optimal one. Bala & Goyal's project provides a point of departure for the investigation into the effects of the amount of connectivity by i) creating a model in which agents use evidence to update their decision making (epistemic) states, and ii) showing that on such a model greater connectivity can lead to less reliability. However their project is concerned with infinite populations, and they examine the communities in regards to economic desiderata. What is needed is explaining what the effects of increased connectivity are for finite populations and in regards to epistemic desiderata.

### 2.3.2 Zollman

Zollman develops a computer model based on Bala & Goyal's approach to investigate the potential instrumental value of epistemic diversity in enhancing a community's reliability in finding the answers to scientific questions. In the model, a finite population of agents is assigned a problem of determining if a new option is better than a known standard, and each agent has a credence concerning the new option's superiority. The new option is in fact better, and it is the job of the scientists to figure this out. Only agents who believe the new option is better will use it and generate new data, and scientists only observe data that they have generated or that has been generated by someone they are connected to on the network. The model can be interpreted as a network of scientists evaluating a new treatment for a disease, where each doctor wants to maximize their patients' survival chances and only administers the new treatment if they believe it is better. Scientists who do not think the new treatment is better rely on their neighbors' research to update their beliefs. Although the medical example is easy to understand, the model can be applied to any scenario where scientists need to assess the success rates of different options to determine their next actions (or tests). Examples include medical researchers in labs, social scientists implementing strategies in schools, as well as market researchers evaluating product strategies.

What is central to each of these examples is that i) scientists update their opinions based on data, and ii) scientists choose which option to research based on that data. As I discuss in the introduction, this latter feature is what I call *preferential research*. It means that scientists generating data for a particular option depends on their opinion of that option. This ends up having a significant role in how Zollman measures the performance of the community. Preferential research means there is a possibility that every scientist abandons researching the new treatment, meaning the community has no chance of ever learning that it is actually better. Zollman gauges the reliability of the community by how often this happens versus everyone in the community reaching a credence of at least 0.9999, what he calls *consensus*. The speed of the community is gauged by how quickly the community is

able to reach consensus in the rounds that they do.

In order to test different levels of connectivity, Zollman is concerned with three network structures. The *complete* network structure is where every agent is connected to every other agent. In this structure, agents hear of every result that is produced each round. The *cyclic* network structure can be easiest thought of as situating the doctors into a ring and having each doctor connect to only their immediate neighbors. In this network, each agent has two connections (neighbors), but no two neighbors have a third neighbor in common (for communities of at least four agents). In this structure, agents only observe a maximum of 3 studies per round (their own and the two on either side). The *wheel* network involves disproportionately distributing the connections so that some members have more influence than others. In this chapter I am focused solely on the effects of the amount of communication, and in Chapter 2 I focus on what happens when that communication is more or less centralized (like it is on the wheel network). For that reason, I focus on cyclic and complete networks and save the analysis of the wheel network for next chapter. (Though it is important to note that results for the wheel network are completely consistent with what is found here.)

To reiterate, there are two possible outcomes for the community in the long-run: they could reach consensus that the new treatment is better (everyone's credence exceeds 0.9999) or they could all stop researching the new treatment (everyone's credence falls to or below 0.5). Zollman measures speed and reliability based on the community achieving one of these two ends.

Zollman shows the complete network gets to consensus faster than the cyclic network (for each size of the population tested). That is, greater communication leads to a faster change in agents' credences. On the other hand, the cyclic network gets to consensus more reliably than the complete network. So complete networks reach consensus more quickly than cyclic ones, but do so less often. For a look at these results, refer to K. J. S. Zollman, 2007.

This represents the Zollman Effect as it has come to be known: First, higher connectivity increases the speed at which the community finds their answer. Second, higher connectivity

decreases the reliability of the answer that they find. In short:

> **Zollman Effect (preferential research version):** Increasing the number of connections increases the speed at which communities change their opinions as well as the chance that communities can unanimously abandon research.

The Zollman Effect can be thought of as a trade-off between speed and reliability that is imposed by the amount of connectivity. Communities can enjoy greater speed through increased connections, but at the cost of reliability. Further, because the greater reliability is achieved by means of a more diverse community, this is often thought to reveal an instrumental value of the diversity. By making the agents more diverse, and ensuring that they continue to test alternatives, the community is more likely to find the optimal option.

### 2.3.3 Similar Results

Similar results can be found in a number of places where philosophers use other mechanisms to cause diversity of opinion in order to stimulate diversity of research. Examples include O'Connor and Gabriel, 2022 and Wu, 2022. The former consider agents with certain levels of intransigence. They show that communities with more stubborn researchers are more able to find the optimal strategy. Similarly, the latter shows that divisions in the community caused by cultural differences can also allow a community a better opportunity to explore its options. The point of both projects (and Zollman's) is to find ways of causing a diversity of opinions because it leads a diversity of research paths. But beyond yielding a diversity of research paths, these projects do not recognize any value to diversity of opinion alone.

## 2.4 Concerns

There are three distinct concerns that one might have about the Zollman Effect.

### 2.4.1 Concern #1: The Zollman Effect Depends on Preferential Research and Does Not Concern Data on Scientific Networks Generally

The first concern that one might have with the Zollman Effect is that it depends on preferential research, and this restricts the applicability of the findings. In each of the models so far, agents only study their preferred theory. This assumption makes sense in the context of doctors that are interested in saving as many patients as possible. But there are plenty of other research situations where agents are interested in learning a particular frequency regardless of how low or high it is. For a few examples: Biologists might want to test the prevalence of a certain species in a region or even a certain feature within a species. Geologists might take samples to determine to what extent a particular substance is present in a particular region. Pollsters might sample what portion of different populations are in support of various propositions. In these situations (and countless others), researchers have an interest in uncovering the true frequency, and their interest is not derivative of maximizing one's expected value elsewhere. So models that assume preferential research limit their applicability to only a subset of sampling-based science. So it is not clear if the Zollman Effect holds for all sampling-based science or a narrow subsection.

Further, even restricted to cases where it does apply, preferential research makes it unclear what is doing the work in increasing predictability. That is, having the agents abandon theories depending on their credence level is a way to *manifest* some of the impacts of increased communication, but it leaves unclear what the impacts are based solely on the change in the amount of information that is exchanged. With the results so far, it is plausible that increased communication is only a detriment to reliability because scientists have the option to abandon research. And so it is not clear if the diversity of their opinions has any use outside of determining what they study. Any value that is found for a diversity of opinions seems dependent on its ability to promote a diversity of research paths. This is a significant concern, because it is already a given that a diversity of research paths is important. Without studying multiple options, we simply would not know about alternatives. So the value of

diverse research paths is not at stake. What is at stake is showing that a diversity of opinions has value. So far, it is not clear that it does without preferential research.

### 2.4.2 Concern #2: The Zollman Effect Only Holds When Parameter Settings Are Tuned Just Right

A second concern one might have does not pertain to the assumptions of the model, but the robustness of the results within the model. Rosenstock et al. (2017) recreate Zollman's model described above and explore a wider range of the parameter space than Zollman considers. They look at greater differences in the performance of either treatment (problem difficulty), greater amounts of tests being taken each round (testing strength), and larger population sizes.

They find that increasing each parameter decreases the difference in reliability between the two networks. Hence there is a less of a benefit to the community's reliability as each of these parameter settings are increased. When the population is big enough, there are enough tests performed per researcher, or the difficulty of the problem is easy enough, then there seems to be little worry that the community will not find the right answer.

On the other hand, the communities get faster with each of these parameter increases. Rosenstock et al. only show this for population size, but a similar story is true for both problem difficulty strength of tests. In these cases, the higher connectivity seems to only have benefit (higher speed) but no downside (no noticeable reliability penalty). Likewise the diversity of the community seems to have no instrumental value. This suggests that the Zollman Effect might only hold for a very particular range of situations, as opposed to it being a general feature of the scientists in this community (whatever the parameter settings). If that were the case, Rosenstock et al. argue, then we should direct our attention and future research to other (more general/robust) dynamics. For a look at these results, refer to Rosenstock et al., 2017.

### 2.4.3 Concern #3: The Result Depends on Bayesianism, and So This Trade-off Is Only an In-house Problem for Bayesians

A third concern that one might be have is that the agents in the model are Bayesian. Bayesianism concerns the particular way in which scientists update their opinions based on evidence, and it also assumes that agents have epistemic states called *credences*. Since the agents are assumed to be Bayesian, one might think that the results demonstrate a feature (perhaps a flaw) of Bayesianism which does not apply to non-Bayesians. Since Bayesianism has not been universally accepted, and one might think that this limits the importance and applicability of the results.

Summarizing the above, there are three distinct worries that one might have concerning the Zollman Effect: i) It might be the case that diverse opinions are useful only insofar as they induce diverse research and have no use otherwise. ii) It might be the case that it only holds for a narrow range of parameter settings, and likewise does not get at any fundamental or principled trade-off between desiderata. iii) It might be the case that this is a problem unique to Bayesian communities and should not worry anyone that does not already accept Bayesianism. These concerns suggest the Zollman Effect is not a fundamental aspect of science so much as a counter-intuitive dynamic for a particular range of situations.

In what follows, I resolve each of these concerns and show how there is a fundamental dynamic inherent to any sample-based science. I do this using two computer models. The first resembles the model that Zollman and Rosenstock et al. analyze discussed above. Using a version of that model without preferential research, I prove a proposition that makes it clear how a version of the Zollman Effect holds for all parameter settings. In doing so I alleviate the first two concerns. The Zollman Effect does not depend on preferential research and is not restricted to any particular region of the parameter space. It is instead a general phenomenon that results from a community consuming more data without producing more data (which is the direct result of greater communication). The second model is a generalization of the

first, dropping even more assumptions. In that model, agents are not Bayesians nor do they have credences. Instead, they only have estimates (which any reasonable epistemology should account for). In doing so, I demonstrate how the Zollman Effect is present in any community where agents update their opinions based on data.

## 2.5 Model 1

### 2.5.1 Setup

Model 1 is in almost all respects identical to the model from above. It concerns a network of scientists that are intent on determining whether a particular frequency is above or below some threshold, and so the interpretation of doctors determining whether a treatment's success rate is above or below that of a known standard still applies. Just like before, each agent has their own opinion concerning how confident they are that the new treatment is better than the old. They individually generate results, share those results, and update their opinions based on all and only the data they acquire. As such, their opinions can be represented by credences (numbers on a 0 to 1 scale). However there are other mathematical representations of the same quantity which I will discuss below. The model should be understood in terms the epistemic state those representations capture —*degrees of confidence*— and not any single representation. With that said, it might prove useful to the reader to think of opinions for the time being as credences, noting that another representation is to come. (Note, the representation of their confidence/opinion does not affect the functioning of the model.)

The single difference between this model and the one from K. J. S. Zollman, 2007 is that scientists will research regardless of what their opinion is. Agents do not have the option to abandon researching the frequency in question, and that means the amount of information that is produced over time is constant.

More formally, the model progresses in successive rounds, and in each round each agent (member of the community) produces a sample of $n$ new data points (patients), out of which

they observe $k$ successes (cured patients). The sample is randomly produced in accordance with the true success rate of the new treatment. Any given study (sample) is likely to have a success rate $\left(\frac{k}{n}\right)$ above or below the true success rate of the treatment, but over time the average of the studies reflects the truth. So for a population of $N$ scientists there will be $N$ many studies and $n \times N$ many individual data points taken (patients observed) each round.

The portion of those studies that any agent sees depends entirely on the network structure. For simplicity I demonstrate my results with the cyclic and the complete networks but it is important to note that the explanations provided depend solely on the amount of connectivity. This is not to say that only the amount of connections makes a difference, since the arrangement of those connections also plays an important role. But the results here demonstrate a *ceteris paribus* law concerning the amount of connectivity.

### 2.5.2 Metrics

Recall that for Zollman, reliability is gauged by how often the community reaches consensus versus collectively abandoning the new treatment. Speed is gauged by how many rounds on average it takes for consensus to be reached (in the rounds that it does occur). In this way he takes measurements for both the speed and reliability of the community, but only infers as to the epistemic diversity of the community. Instead, I find useful to gauge the epistemic diversity directly, and also gauge the community over time.

I use three primary metrics concerning agents' opinions, i.e. how confident they are that the new treatment is more successful than the old. *Diversity* is the variance between the agents' opinions. While the philosophical notion of epistemic diversity should not be identified with any single metric, this one has the intuitive feature that greater variance corresponds to a greater spread in the agents' opinions. Maximal variance would mean the community is evenly split between either extreme. This form of diversity should not be confused with other forms, like the diversity of research paths. That form of diversity concerns the difference in what theories are being tested. In this model, all agents test the same theory every round, so there is no diversity of research paths. However there is diversity

of opinions when agents hear of different pools of evidence.[2]

The other two metrics pertain to the aggregate viewpoint of the community. By *simulation*, I mean a specific configuration of parameters. An individual execution of the model using those parameters for a specified number of rounds is referred to as a simulation *run* or *iteration*. The term *aggregate opinion* denotes the average degree of certainty in one of these simulation runs. The expected aggregate opinion for a simulation is given by the average of the aggregate opinions for the runs of that simulation. By *speed*, I am referring to the expected pace at which the aggregate opinion changes. The *predictability* of the community is gauged by the variance between the aggregate opinions of different runs of the same simulation.

It is important to notice the relationship between Zollman's and my metrics, especially with regard to how the reliability of the community is captured. The crucial point is this: the more unpredictable a community is in my model, the more likely that community would be to abandon research had they been preferential researchers. Similarly, the more likely a community would be to abandon research (if they were preferential researchers), the more unpredictable their aggregate opinion will be if they are not preferential researchers. The same is true for speed as well: the faster a community is in my sense, the quicker they will reach consensus in Zollman's (and vice versa). And while Zollman does not directly measure diversity of opinions (aside from whether the community is all above or below some threshold), the diversity that he discusses finding instrumental value of (diversity of opinions) is the same that is captured here.

With these metrics in mind, it is helpful to clarify what the Zollman Effect is in these terms and likewise what the goal is. I will show that higher connectivity increases the community's speed, decreases its diversity, and as a result decreases its predictability. This will show not only a trade-off between the speed and predictability, but will also show a usefulness of diversity (that it increases predictability).

[2]There is also a diversity of opinions due to an initial distribution of the agents' confidence levels, but this is quickly overpowered based on whatever their respective pools of evidence dictate.

Before turning to the results of the model, it turns out to be helpful to consider another representation for the agents' opinions in addition to credence. A credence represents an agent's level of confidence on the $[0, 1]$ scale. A logarithmic scale, which I call *logodds*, allows us to describe epistemic states more conveniently (and accurately too — since accuracy is often sacrificed for the sake of convenience). It can be obtained by performing a logarithmic transformation on an agent's credence using the following formula:

$$logodd = \log_{10}\left(\frac{credence}{1 - credence}\right)$$

The transformation is easiest understood with examples such as the following:

| Credence | | Logodd |
|---|---|---|
| | $\vdots$ | |
| .999 | $\rightarrow$ | 3 |
| .99 | $\rightarrow$ | 2 |
| .9 | $\rightarrow$ | 1 |
| .5 | $\rightarrow$ | 0 |
| .1 | $\rightarrow$ | -1 |
| .01 | $\rightarrow$ | -2 |
| .001 | $\rightarrow$ | -3 |
| | $\vdots$ | |

Note that an agent's logodd is essentially a count of the number of 9's after the decimal place (or 0's followed by a 1).

Since logodds are simply another representation of the same quantity that credences are, using them does not depend on any further philosophical assumptions about the community (how they are setup or how they perform). Instead, they merely offer another way of *analyzing* the members. This alternate lens turns out to be very helpful, because it makes the Zollman effect abundantly clear.

One might worry that this makes the above definitions concerning the aggregate and

variance of opinions ambiguous between referring to credences or logodds. That is done on purpose because the relevant analysis for the Zollman effect holds for both. The speed at which the logodd changes is directly correlated with that of credences, and the same holds for the predictability. So if the trade-off can be found for logodds, it holds for credences as well. More accurately, the trade-off concerns agents' *confidences*, of which credences and logodds are merely two different mathematical representations. Neither should be confused with being *identical* to confidences, nor should one have a monopoly concerning how we analyze them.

### 2.5.3 Results

In this section, I show how the following is true for Model 1:

> **Zollman Effect (general version):** Increasing the amount of connectivity increases the rate at which community aggregate opinion changes, makes the changes of that aggregate opinion less predictable, and reduces the overall amount of diversity between the individual opinions.

The following proposition says that the logodd an agent can expect to have is a linear function of the amount of data that they acquire.

$$\mathbb{E}[logodd_{new}] = logodd_{old} + n \times B$$

$\mathbb{E}[]$ indicates the expected value[3] of a random variable, $n$ is the number of trials that an agent has observed since their last update (to get to $logodd_{old}$), and $B = 2\epsilon \log_{10}\left(\frac{5+\epsilon}{5-\epsilon}\right)$ which is constant for the simulation. $B$ depends on $\epsilon$ which is the difference between the success rates of the treatments, and so dictates the difficulty of the problem at hand. (A smaller $\epsilon$ corresponds to a more difficult problem, because they are harder to distinguish.) To put it

[3]The expected value of a random variable is also known as its mean. That is, the average value of this variable is its expected value. In this case, the expected/average value of the new logodd means the value that the logodd will be on average (if the same process was simulated infinitely many times).

simply, the proposition states that the logodd that one can expect to have is a linear function of the amount of evidence that they observe in proportion to the difficulty of the problem (how close the success rates of the treatments are). So more evidence for an agent means a greater expected logodd. (Note a similar proposition can be proved in terms of credences.)

(One might worry that the agents knowing the value of $\epsilon$ makes the model unrelialistic in that it makes the agents too knowledgable about the problem. In real world cases, doctors do not know in advance exactly how much better or worse the new treatment will be compared to the old. However K. J. S. Zollman, 2010 has a model that drops this assumption, so I do not take it to be a live concern that needs to be addressed like the other three I list. And even if it did need to be addressed, model 2 drops the assumption also. Assuming that agents know the value of $\epsilon$ is an innocent idealization, like physicists assuming a frictionless surface to focus on particular dynamics of interest. With it assumed, the dynamics become clearer, and proving results analytically is more feasible (like I do above). This directs our attention and teaches us how to navigate the more complicated models. The particular dynamics I am concerned with pertain to how statistically generated data is used in communities to form opinions.)

The proposition seems intuitively obvious once understood, and likewise its implications can be easily overlooked. Specifically, the Zollman Effect itself follows directly from it. Recall that the Zollman Effect says that greater connectivity leads to a faster, less diverse, and less predictable community, while less connectivity leads to a slower, more diverse, and more predictable community. The rest of this section explains why.

Since greater connectivity means that each agent acquires more studies each round, proposition 1 entails that their expected logodd will be higher. And since this will be true for everyone in the community, the aggregate logodd of the community will be higher. So a highly connected community is able to reach higher aggregate confidences than a less connected community using the same amount of data. This is reflected in Figure 3.1 which shows the average aggregate logodds for cyclic and complete networks over time. Notice

that each of the networks' aggregate logodds are linear functions of rounds passed. Each community can expect the same change from one round to the next, however that change is larger depending on how many connections there are. Hence, greater connectivity causes higher speed.

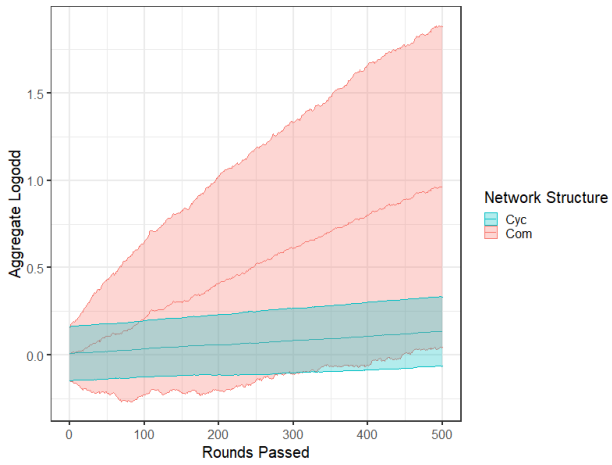



Figure 2.1 A demonstration of the Generalized Zollman Effect. Notice that while the more highly connected network increases confidence (logodd) more quickly on average, any given run is much more difficult to predict than on the cyclic network. ($\epsilon = .0005$, $n = 100$)

Next, consider how connectivity influences the overlap between agents' pools of evidence. If there were no connections, there would be no overlap between the data that any two agents see. Every connection added to the network makes it so the connected agents see each other's data, and so there is greater overlap in their pools of evidence. But if two agents have different neighbors, then their pools of evidence will have at least some difference based on what they hear from their respective neighbors. When everyone in the network is connected, there is complete overlap between everyone's individual pools of evidence. Hence greater connectivity decreases the diversity of the community. This can be observed in Figure 2.2. Notice that for the cyclic network there is an initial increase in diversity. This is because the studies are statistically generated with their own variance. When the agents spread, they are reflecting this variance. Over time, everyone's evidence leads them to approach credence 1 (which is

a consequence of proposition 1), and so the variance of their credences disappears. However the more the network is connected, the less they will spread initially and the quicker the variance will disappear. This is because the amount of overlap between the agents limits the potential for them to have diverse evidence. Since the diversity of opinions depends on the diversity of evidence, it is curtailed too. Hence, greater connectivity causes lower diversity.

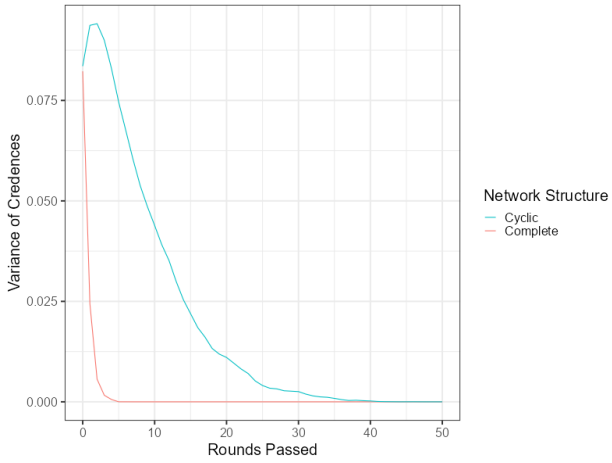



Figure 2.2 The average variance of credences over time ($\epsilon = .05$, $n = 10$)

To see that higher connectivity makes the community less predictable, we can make two comparisons: First, consider how much evidence cyclic and complete networks are expected to require in order to achieve a particular aggregate logodd (i.e. y-axis value from Figure 3.1). It takes the cyclic network longer to reach the same point, because (as discussed above) the cyclic network is slower. But this means that the cyclic network will have produced (and its aggregate opinion will reflect) many more results than the complete network. Because of this, it will be more likely that the aggregate opinion of each iteration of a cyclic networks stays close to what it is expected to be. On the other hand, the complete network iterations are expected to reach the same aggregate using less information. Hence the greater connectivity makes it more likely that the aggregate on any given run will be farther from what it is expected to be — in both the positive and negative directions. This can be seen by

the standard deviations shown in Figure 3.1. Notice that for any two points from either network that sit on the same horizontal line (y-value), the standard deviation of the complete network is greater than that of the cyclic network. So with regard to any particular expected aggregate logodd (or credence) when a network is more connected, it is less likely that any given iteration will be close to that expected aggregate. In other words, higher connectivity makes networks less predictable.

The other comparison we can make concerns networks of different levels of connectivity at the same moment in time, and likewise, based on the same amount of results produced (x-axis value in Figure 3.1). Recall that diversity is caused because the studies themselves are generated with their own variance. Typically the results of all the agents will generally be close to the true success rate, but sometimes, some of the agents will produce studies that indicate a success rate far above or below what the true one is. In these cases, the level of connectivity plays a significant role. Only the agents that produce the results as well as their neighbors will update on the unlikely results. The more neighbors that they have means more agents will update on the studies. This in turn means more of the community will be affected by the unlikely results. Hence the aggregate community opinion will likewise be more skewed by these particularly different results. Conversely, with less connectivity the community members are insulated against the results. While some members will change their credences significantly, the rest will do so based on the (typical) results that they have produced. This means higher connectivity leads to the community being less predictable on the same amount of evidence produced.

To summarize, higher connectivity makes a community faster, but in doing so it makes it less diverse and likewise less predictable. This captures the same trade-off that Zollman finds: greater speed can be achieved by sacrificing predictability. Additionally, this demonstrates the usefulness of diversity that is typically associated with his result: Greater diversity leads to greater predictability. This usefulness does not depend on it promoting a diversity of research paths, but purely based on how it reflects the diversity of the evidence itself. This

shows that preferential research is not necessary for diversity of opinions to be valuable.

Since preferential research did not play a role, these results alleviate concern #2 — that the Zollman Effect depends on the possibility that agents might abandon research. Further, notice that none of this was relative to any particular parameter settings. The proposition is true for all values of $n$ (sizes of the samples taken each round), all sizes of the community, and all problem difficulties. This alleviates concern #1 that the Zollman Effect is not robust across parameter settings and is merely a feature of a niche combination of parameter settings. This shows instead that there is a more general effect underlying what Zollman discovers, and this effect concerns the fact that greater connectivity leads to consuming more information without producing more.

### 2.5.4 Possible Objections

#### *2.5.4.1 Following Up with Rosenstock et al.*

One might worry that this does not alleviate the heart of the concern that Rosenstock et al. raise. They might contest that Rosenstock et al. show that the effect becomes insignificant for some parameter settings (albeit maybe present in a minimal way), and that is why it does not deserve further attention.

It is worth pointing out that Rosenstock et al. show the effect diminishes as conditions get *better* (stronger tests, more researchers, greater difference in treatment success rates). Rephrasing their results, we can say that as conditions get worse (smaller sample sizes, fewer testers, closer success rates), the Zollman Effect gets more significant. That makes the Zollman Effect *more* interesting, because it is the more challenging cases that we should be interested in in the first place. The fact that the trade-off matters less when the challenge itself matters less does not make the trade-off less interesting.

Setting that aside, Rosenstock et al. are right to show how the degree of this effect relates to the various settings. It is important to keep this in mind when trying to measure for these results or recreate findings, because in many situations the metrics might not be calibrated

appropriately for the particular parameter combination. But this does not change the fact that there is a fundamental dynamic in play which the results here expose. That dynamic deserves our attention, because it is a fundamental aspect of science. Even if its effects might not seem interesting in such a simplified model, they might have extreme impacts in more complicated environments. It is important to remember that this exploration is about discovering *fundamental* dynamics.

If we focus too closely on specific values of variables, or on any particular representation which might not fully capture the model, then we risk missing the point of this methodology. Models are not meant to be tuned to particular real-world examples. They are meant to unfold dynamics that are otherwise obfuscated by the complexities of the real world.

#### *2.5.4.2 Different Instances of Value?*

A different objection that one might make concerns whether or not my mechanic is truly capturing a deeper cause of the unreliability found in Zollman and elsewhere. One might concede that this demonstrates some sort of usefulness for diversity of opinions, but insist diversity of research paths have their own importance. Likewise, they might reject that these results somehow explain Zollman's, because his get at something else that is valuable.

The problem with this objection is that it misunderstands the significance of Zollman's results. There was never any doubt that a diversity of research plans is important. His model (nor any of the others in the literature) does not simply demonstrate that communities of agents that research diversely will perform better. We know that a diversity of research is valuable because you simply cannot discover certain things without it. Zollman builds a model where a diversity of opinion is needed to obtain a diversity of research, and then identifies a factor that leads to a diversity of opinion. I suggest that even without this mechanism, a diversity of opinion leads to predictability across runs, and that is the *real* import of the Zollman Effect.

So I fully concede that there is a value to diversity of research paths independent of what I show here. But I deny that Zollman's results discover an importance of that sort of diversity.

They take for granted the value of that diversity (as is perfectly reasonable) and attempt to find causes for it. The same is true for the rest of the models in the literature. They are concerned with finding causes for a diversity of opinions that in turn yields a diversity of research paths. The point of interest for each of them is the role that diversity of opinions plays, namely, that it can promote a diversity of research paths.

This project shows the impact of the diversity of opinions on how predictable the community is, without any impact on the research of the community itself. Further, as mentioned above, a community being less predictable in my sense directly corresponds to the possibility that it abandons research when preferential attachment is in play. The high unpredictability means that there are more cases where communities whose aggregates are low, reflecting a set of members all with low opinions as well. It is in these cases that research is abandoned had they been preferential. Therefore, not only is the dynamic I discover more general, it also gets at the root cause of Zollman's which preferential research merely manifests.

To summarize: in previous results diversity of opinion is found to be useful because it is able to promote diversity of research paths. But the reason why it is able to promote a diversity of research paths is because it properly appreciates the diversity of the studies produced. And I have shown that doing so has upside regardless of whether or not it also impacts diversity of research paths.

## 2.6 Model 2

### 2.6.1 Setup

The third and final concern that I consider is that the Zollman Effect is only a Bayesian issue, since the above model (both Zollman's version and mine) have the agents using Bayes' rule. This final model is meant to alleviate this worry by dropping the Bayesian assumption. But before getting into it, it is worthwhile to note the following: While the above is consistent with Bayesianism, nothing in the explanations of the results relied on that update strategy versus another. Therefore, it is already doubtful that Bayesianism is playing any significant

role (beyond what any plausible update strategy would play). But for anyone that is still suspicious, the following model will put this concern to rest.

Model 2 also drops the assumption that agents have *credences* (or logodds). Those particularly opposed to formal epistemology might think of these results to be a consequence of the feature of Bayesianism to require its agents to have credences. But the results that follow depend on only the agents updating their *estimates* which any plausible epistemology must account for.

Model 2, similar to model 1, involves a community of scientists intent on finding the best treatment, except there are multiple treatments to evaluate and the agents have no prior knowledge about any of them. This model is similar to a model from Kummerfeld and Zollman, 2016 in that agents have more than one unknown option open to them, and they do not know any extra information about the success rates of those options (like the value of $\epsilon$ in the last model). However the agents on Kummerfeld & Zollman's model have credences and update those credences using Bayes' rule. So their model does not alleviate concern #3. Instead of capturing the agents' epistemic states by their credences, I represent them by their *estimates*. An estimate is nothing more than what the agent thinks the true frequency actually is. The agent updates their estimate in as innocent a way as can be: an agent's estimate for a treatment's success rate is simply the amount of successes they have observed divided by the total amount of patients. In other words, as agents acquire data, they simply change their estimates to be exactly what their data (over time) says. By just using estimates and data, this model requires minimal epistemological assumptions. Even so, the Zollman Effect obtains.

The agents each produce evidence for every option every round (they do not prefer some option over others based on their estimates - which would be a version of preferential research). So there is a constant amount of evidence being generated each round for every option. The agents share all of their results with their network connections each round.

### 2.6.2 Metrics

Since the agents do not have degrees of confidence (credences/logodds), performance on model 2 cannot be gauged in the same way as model 1, so new metrics are needed. As I have mentioned above, the Zollman Effect should not be *identified* with any particular metric (even with respect to a particular model). Instead, it can be captured by different metrics in better and worse ways. For this model, I choose a very simple way of demonstrating the Zollman effect which concerns agents' abilities to find the best option (the treatment with the highest success rate). An agent has found the best option if their estimate for what is in fact the best option is higher than their estimate for any other options. That is, they have properly identified the best option as the best one. This is what it means to say that an agent is *correct*.

The speed of the networks can be gauged by the change in the average amount of correct agents over time for either network. It is worth noting that nothing I have said guarantees a community's speed stays constant (unlike what proposition 1 guarantees for model 1), so the speed of the networks might change over time. It turns out that the change of speed is one way to gauge the predictability, or at least stability, of the networks.

Another way to gauge the predictability of the community involves ordering the iteration results for each network by the average amount of correct agents throughout the simulation (starting from the iterations with the most amount of correct agents and ending with the iterations with the fewest). Next, the two ordered lists (one for each network) are broken into three brackets (for each network): the top 25%, the middle 50%, and the bottom 25%. So the top 25% of the cyclic iterations are the quarter of cyclic iterations that had the most amount of correct agents (compared to the rest of the cyclic iterations). The bottom 25% are the iterations with the fewest amount correct agents.

Using these six brackets (three per network), the predictability of the networks can be gauged by plotting the average performance (number of correct agents) of each bracket against the number of rounds passed. The predictability is gauged, not by standard devia-

tion, but by how closely the brackets for a network are clustered. The closer these brackets are together, the more predictable a community (network) can be said to be. A smaller difference between the top and bottom brackets indicates that on average there is less of a difference in the amount of agents finding the correct answer in the best and worst scenarios. A greater separation between the brackets means that any given run is less likely to be close to the average of all the network's iterations — it is less predictable what that run will do.

### 2.6.3 Results

Figure 2.3 shows the average performance of the networks without breaking them up into brackets. As expected, time passing (and evidence accumulating) allows more agents to correctly identify the best option for both networks. What is important is that the complete network is on average faster at achieving a higher amount of correct agents. In each round, more agents on average find the correct answer. This is because agents get more data for every option each round, and the law of large numbers dictates that their estimates will necessarily get more accurate. Hence they will be in an increasingly better position to identify the best option *earlier* than they would be able to on a cyclic network (with less information).

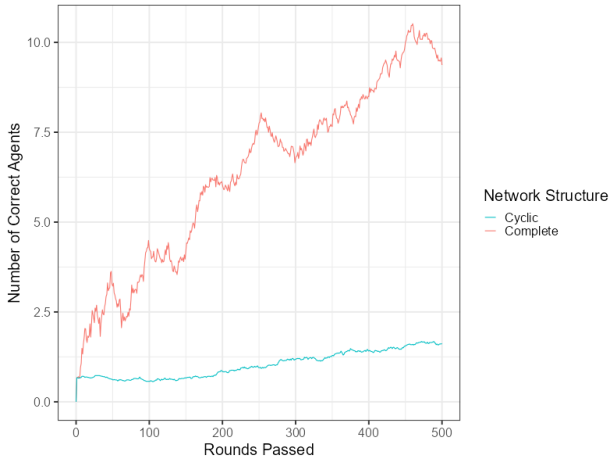



Figure 2.3 The amount of doctors (out of 24) that find the correct theory (out of 64).

Figure 2.3 also shows that performance on the complete networks is in some way less stable, i.e. predictable. A different set of iterations will have peaks and valleys in different locations. Compared to the cyclic network, one is in less of a position to guess how many correct agents any given run of a complete network will have. This is for the same reasons as above: The greater communication makes it so there is less overlap in the pools of evidence. This makes it more likely that larger portions of the community will either succeed or fail together. They jump from one favorite to another together.

Figure 2.4 shows the amount of correct agents for each bracket of the networks. Notice how closely grouped the three brackets of the cyclic network are in comparison to how spread out the performance of the complete network is. The difference between the best and worst on a cyclic network is much smaller than it is for the complete network. As explained above, this means that it is more predictable what any given run of the cyclic network will do (versus a complete network). Notice the top 25% as well as the middle 50% of the complete are higher than the top 25% of the cyclic, meaning that the complete network has higher potential in the best cases. However also notice that the bottom 25% of the complete are lower than the bottom 25% of the cyclic. So in cases where ensuring that at least someone finds the truth is important, choosing the more highly connected option would be counterproductive. These features show that the more connected networks are less predictable than the cyclic networks.

One might think that the higher number of correct agents is worth the risk. But in those cases, the agents are essentially lucky to find the truth. They could just as easily have had zero agents finding the truth since just as many runs ended that way. This makes it questionable whether it is appropriate to say the agents on the complete network, even in scenarios where more of them are correct, are in a better epistemic state. It might be that they have a more questionable *justificatory* status. That is, their beliefs are not *safe* in the sense that they could have easily been wrong.

Both figures demonstrate the same underlying dynamic: the Zollman Effect. The greater

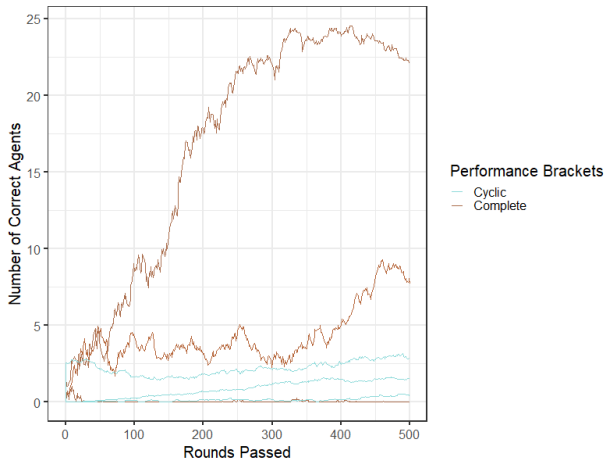



Figure 2.4 The amount of doctors (out of 24) that find the correct theory (out of 64). The performance of each network is broken into three brackets. The highest performing bracket represents the average amount correct agents for the top 25% of runs of the simulation. The lowest bracket shows the average amount of correct agents for the 25% of runs with the lowest amount of correct agents throughout the simulation. The middle bracket shows the average performance of the other half of runs for a simulation which are neither in the top or bottom 25%.

connectivity of the complete network leads to it having a quicker average increase in the amount of correct agents. However this also leads to the community being less diverse. More connections means each agent has more information, but there is less of a difference between their information and that of those around them. This means the community is at greater risk of a majority preferring a subpar option. When there are fewer connections and likewise greater diversity between the agents' pools of evidence, majorities usually only form when in support of the optimal option.

This demonstrates the Zollman effect in a way that requires minimal assumptions. The agents are not assumed to be Bayesians nor have credences, they are not preferential researchers, and nothing in the above explanation hinged on the particular parameter settings — aside from the network structure. This shows that the amount of connections in a community imposes a trade-off between being fast and being predictable. Similarly, this demonstrates that epistemic diversity can help a community achieve higher predictability

regardless of its impact on research paths.

## 2.7 Conclusion

I have shown the amount of connections imposes a trade-off for communities that use probabilistically generated data to update their epistemic states. In particular, greater communication allows for a better *average* epistemic state sooner, but the variance between the possible epistemic states is greater. This in turn demonstrates a usefulness of epistemic diversity in its ability to curtail the unpredictability of a network. Further, none of this depended on any particular parameter settings, preferential research, or commitments to Bayesianism.

So what is the upshot? First, it shows an extremely counter-intuitive result, the Zollman Effect, is actually a general phenomena for any communities that update their epistemic states based on evidence. Having more data is not necessarily better. Intuition suggests (and the law of large numbers supports) that more data can be nothing but good for an agent. This project shows that while producing more data may be beneficial, the same is not necessarily true about how much is shared. This demonstrates that sharing data can inflate/distort/or otherwise impact the effects of data on a network. How one shares data (even just numbers without analysis) dictates how the community performs epistemically.

Second, it informs the debate concerning the value of diversity. Diversity is typically thought of as a negative feature of communities. The classic example of diversity is polarization, which is thought of as a bad thing. Zollman's original result helped open the door up to the possibility that diversity can sometimes be valuable. However one can achieve Zollman's ends without diverse opinions (by causing diverse research in some other way). Therefore one might be skeptical based on his results that diversity of opinions really plays a significant roll, or if it is right to attribute value to it outside of its ability to promote research paths (which we already know are valuable). This project alleviates that worry. The increases in predictability discussed above depend specifically on agents having different opinions.

A further aspect of these upshots is that what might be best for an individual is not

necessarily best for their community. Making individuals more predictable is done by is done by giving them as much evidence as possible. However this leads to the community's opinion being *less* predictable.[4] Therefore a community having agents that are more predictable can cause the community itself to be less predictable. This concerns what Mayo-Wilson et al., 2011 calls the *independence thesis*: that achieving the epistemic desiderata for communities is independent of achieving them for its members. Members consuming more information in their community might be best for that member, but not the community as a whole. Now, *both* the community and the individual are made more predictable if the agent *generates* more data, but the question is whether or not an agent should get more of the data that is *already* present in a community. This project shows that it can be in the community's best interest (though not necessarily the individual's) to impose a division of cognitive labor by dividing who updates on what evidence.

### 2.8 Additional Material

A direct consequence of Bayes' rule is that:

$$\frac{P(H|E)}{P(\neg H|E)} = \frac{P(H)P(E|H)}{P(\neg H)P(E|\neg H)} \tag{2.1}$$

$$= \frac{P(H)}{P(\neg H)} \times \frac{P(E|H)}{P(E|\neg H)} \tag{2.2}$$

Note, $\frac{P(H)}{P(\neg H)}$ are the *prior odds* that the new treatment is actually better, and $\frac{P(H|E)}{P(\neg H|E)}$ are the *posterior odds*. Taking the logarithm of both sides yields the formula for updating from one's prior to posterior *logodds*:

[4] I am not comparing the community to the individuals here, but the community to itself in either case (and similarly for individuals).

$$\log_{10}\left(\frac{P(H|E)}{P(\neg H|E)}\right) = \log_{10}\left(\frac{P(H)}{P(\neg H)} \times \frac{P(E|H)}{P(E|\neg H)}\right) \tag{2.3}$$

$$= \log_{10}\left(\frac{P(H)}{P(\neg H)}\right) + \log_{10}\left(\frac{P(E|H)}{P(E|\neg H)}\right) \tag{2.4}$$

In other words:

$$l_{t+1} = l_t + \log_{10}\left(\frac{P(E|H)}{P(E|\neg H)}\right) \tag{2.5}$$

Since each observation is an independent reflection of the frequency in question, the probabilities that the results they see are caused by either of the two possible frequencies ($0.5 + \epsilon$ or $0.5 - \epsilon$) are given by the following:

$$P(E|H) = \binom{n}{k}(.5+\epsilon)^k(.5-\epsilon)^{n-k} \tag{2.6}$$

$$P(E|\neg H) = \binom{n}{k}(.5-\epsilon)^k(.5+\epsilon)^{n-k} \tag{2.7}$$

And so:

$$\frac{P(E|H)}{P(E|\neg H)} = \frac{\binom{n}{k}(.5+\epsilon)^k(.5-\epsilon)^{n-k}}{\binom{n}{k}(.5-\epsilon)^k(.5+\epsilon)^{n-k}} \tag{2.8}$$

$$= \frac{(.5+\epsilon)^{2k-n}}{(.5-\epsilon)^{2k-n}} \tag{2.9}$$

$$= \left(\frac{.5+\epsilon}{.5-\epsilon}\right)^{2k-n} \tag{2.10}$$

Therefore:

$$l_{t+1} = l_t + \log_{10}\left(\left(\frac{\epsilon + .5}{\epsilon - .5}\right)^{2k-n}\right) \tag{2.11}$$

Since the frequency is actually $0.5 + \epsilon$ (the new treatment is better), the expected value of $k$ is $(\epsilon + .5)n$, therefore:

$$E[l_{t+1}] = l_t + 2n\epsilon \log_{10}\left(\frac{\epsilon + .5}{\epsilon - .5}\right) \tag{2.12}$$

# 3. WHEN DIVERSIFYING OPINIONS IS USEFUL

## 3.1 Overview

This chapter proposes a mechanism that explains how diversifying opinions can be useful. Two hypotheses are tested and rejected, leading to the development of a third hypothesis that identifies the appreciation of diverse evidence as the key mechanism driving the advantages of epistemic diversity. I show that diversifying a community's opinions can make its aggregate opinion more *predictable* when it is done in a way that reflects the diversity of the *evidence* that is produced.

## 3.2 Introduction

In certain scientific communities, epistemic diversity has been argued to have instrumental value in certain cases, however not everything which causes diversity of opinions is beneficial. In this chapter, I aim to identify the mechanism that underlies the usefulness of epistemic diversity by falsifying two hypotheses. Ultimately, I propose a third hypothesis that the usefulness of diverse opinions stems from reflecting whatever diversity there is of evidence produced.

I do this by building on Chapter 1 where I show a particular case in which diversifying opinions is useful. In that chapter, I use computer models that simulate scientific research to show that fewer connections leads to greater diversity, and as a result, higher predictability. In this chapter, I consider further complications/generalizations of one of the models to pin down why (and when) epistemic diversity can be useful. In particular, I consider agents that have various levels of precision (some are better at generating data than others), different distributions of the network connections (some agents agents communicate more than others), as well as a form of intransigence (where agents weigh their own evidence more significantly than the rest). Analyzing the effects of these changes reveals the mechanism responsible for when diversity does lead to predictability: Diversifying (increasing the variance of) opinions leads to predictability when the the result is a distribution of opinions that is closer to (has a more similar variance to) that of the evidence that is generated by the members.

This contributes to a growing discussion concerning the value of epistemic diversity. Many philosophers have argued for an instrumental value in diversity of opinions due to its ability to promote a diversity of research paths. In particular, K. J. S. Zollman, 2007 establishes what has been since called the *Zollman Effect*, which is that decreasing connections, and thereby increasing diversity, can make a community less likely to abandon optimal research paths. In Chapter 1, I generalize the Zollman Effect. I show that there is a more fundamental usefulness to the diversity of opinions that explains *why*, at least in the case of decreasing connectivity, it is able to increase a diversity of research paths. It is because doing so allows

the community as a whole to better reflect —at a societal level— the evidence produced. That is, less connectivity means that the distribution of the community members' opinions has a closer variance to the distribution of the evidence that is produced. In this way, the community aggregate reflects the diversity of the evidence in a way that it does not when the members all have the same opinions.

In Chapter 1, I reject the necessary condition that for diversity of opinions to be beneficial, it must be to promote a diversity of research paths by considering a model in which all agents research no matter what. In this chapter, I defend a sufficient condition for when promoting a diversity of opinions is beneficial: when it reflects the diversity of the evidence produced. But before reaching this conclusion, I first falsify two less nuanced hypotheses. It is instructive to do so, because it highlights cases when diversity of opinions is not beneficial and as result, the mechanism at play when it is.

This chapter proceeds in three main sections. Each involves its own complication of the model that impacts the level of the diversity of the agents. By showing which ones are not beneficial and which are, I identify the mechanism at play for when diversity of opinions is beneficial.

## 3.3 Preliminaries

### 3.3.1 Value of Diversity Thus Far

A helpful survey of the literature on the value of epistemic diversity can be found in O'Connor and Wu, 2021. In that article, transient diversity is defined as the temporary period during which agents research competing strategies. Beginning with Kuhn, 1977, Kitcher, 1990, and Strevens, 2003, philosophers have argued in various ways for a value in having different researchers test different theories. A theme between these projects is that the scientific community is able to more efficiently explore the epistemic space when research paths are diversified.

This theme is continued in the network epistemology literature where several projects

focus on promoting diversity in research. For example, Kummerfeld and Zollman (2016) examine the role of connectivity in promoting diverse research paths, while O'Connor and Gabriel (2022) investigate the impact of bias, and Wu (2022) focuses on cultural divisions. However, these philosophers only consider the value of generating a diversity of research, and do not recognize any value in diversity of opinion aside from that. The agents in each of these models research only contingently, when their opinions dictate it. This is what I call *preferential research*. This suggests that promoting diversity of opinion is only valuable insofar as it leads to a diversity of research paths, and may not be useful when researchers are already exploring different avenues of research.

In Chapter 1, I show that there is a usefulness to diversity of opinion even without preferential research. That is, even when opinions do not make a difference to what is researched, it still is beneficial for a community's predictability that they be diverse. I discover a fundamental dynamic that concerns the flow of information alone and explains Zollman's findings. It turns out Zollman's inclusion of preferential research *manifests* what I discover, but there is an important dynamic even when it is excluded.

Chapter 1 is concerned with epistemic diversity as a function of the level of connectivity, and it shows that diversifying opinions as a result of that particular change can be useful even without preferential research. This chapter considers the usefulness of epistemic diversity more generally. I determine when diversity of opinions is valuable when caused by things other than the amount of connectivity.

### 3.3.2 Base Model

Each of the following sections considers a single complication to a particular base model, which is one of the models I use to analyze scientific communities in Chapter 1. The community consists of $N$ many doctors that each produce $n$ new data points every round. The observed frequency of a study is $\frac{k}{n}$ where $k$ denotes the number of successes for the study. The $\frac{k}{n}$ of any given study is likely above or below the true frequency, but overtime the aggregate frequency of the studies reflects it. As mentioned above, the most important difference

between my models and others in the literature[1] is that in theirs, agents' research is contingent. On my model, agents research every round, and so this entails the amount of evidence that is produced each round is constant. This ensures that the dynamics that unfold are due purely to the flow of information, and not how or if it is generated.

Every round $N \times n$ many data points are produced in the community, but the amount of those data points that any given member of the community observes is dependent on the network structure. Agents only observe the studies of those they are connected to on the network. Chapter 1 concerns how the amount of connections on a network affects its performance. In the next section I explain these results, and show how they demonstrate at least one way in which inducing diversity can be useful. But first, it is helpful to see three reasons why the model provides a suitable environment to launch this investigation into the usefulness of diversity of opinions.

First, it captures an integral part of a significant span of scientific research which concerns updating opinions based on data. The dynamics discovered here are not relative to only trivial problem solving games but the *general* features for how statistical data is shared in communities. This is a core aspect of scientific research, and one where one might easily think dynamics are innocent or trivial. It is easy to think more data is always better for achieving any and all epistemic desiderata, but the results of this model (both here and in Chapter 1) demonstrate that is not always correct.

Second, while other models in the literature have focused almost exclusively on the average performance of communities, it is important to also consider variance measures. These measures give insights into diversity and predictability that have so far been ignored. By taking the variance between a community's aggregate opinions on different runs of a simulation, we can gauge the predictability of the community. On the other hand, by taking the variance of the individual opinions within the community during any given run, we get a measure of the spread of the community members' opinions. The average amount of variance

[1] and the single difference with regard to one of the models used in K. J. S. Zollman, 2007, 2010

across all runs is the average amount of diversity for that simulation.

Predictability is important because it affects the reliability of a community's opinion. Even if one doesn't take the idea of group epistemic desiderata seriously in their own right, there are countless situations where agents depend on the performance of the group. For instance, if agents have to vote for what they think is the right answer or if they are preferential researchers, the reliability of their opinion is tied to their predictability. In this way, even if the community itself doesn't have epistemic desiderata, there are likely many cases in which individuals still have reason to be concerned with the findings on predictability.

Third, this model allows us to distinguish between a number of different types of epistemic diversity and articulate the relationships between them. *Diversity of opinions* refers to what is discussed two paragraphs above: the spread between agents' individual levels of confidence. *Diversity of evidence held* refers to the disparity between the evidence that agents hold, i.e. to what extent there is an overlap in the data each update their respective opinions on. This is defined in terms of there being any correlation between two agents' pools of evidence. When they do not share evidence between each other, their pools of evidence are entirely independent from one another (though based on identical distributions). When agents share their evidence, there is a correlation between the studies they observe and likewise a correlation between their opinions. *Diversity of evidence produced* refers to the diversity of the evidence produced by the entire community, regardless of who observes it. That is, the process of random sampling has its own inherent variance which this reflects. (The existence of this type of variance is why we take multiple samples.) Decoupling these notions and articulating the relationships between them turns out to be the vital to understanding the ways in which diversity can be beneficial.

I represent agents' opinions (levels of confidence) in two ways: on a 0 to 1 scale, normally referred to as *credences*, and on a logarithmic transformation of that scale from $-\infty$ to $\infty$, which I refer to as *logodds*. As these are merely different metrics of the same quantity, they do not require any further assumptions. In future work I plan to focus specifically on the

subtleties of using either metric. Here it is sufficient to recognize that both give legitimate perspectives on the behavior of the community. Showing that diversifying credences curtails the unpredictability of logodds (like I do in Chapter 1) shows that there is some sense in which making a community's opinions more diverse makes them more predictable. Seeing how this happens turns out to be very instructive, because it focuses our attention on how evidence is distributed and what it might mean for a community *as a whole* to over count studies. The particular metrics facilitate explaining the mechanic, but they do not affect the mechanic itself. At the end of Chapter 1, I show a more general model that does not use either metric and shows the same effect occurs with estimates alone (which any reasonable epistemology must account for). The same is possible here. But using credences and logodds is simply more helpful for explaining and understanding the dynamics.

### 3.4 Network Connectivity

In Chapter 1, I show that decreasing the amount of connections on a network can make the community more predictable. A more detailed analysis of the impacts of connectivity can be found in that chapter, but here only a brief recapitulation is necessary. Figure 3.1 displays the average and standard deviation of the aggregate logodds of two different networks. The *complete* network represents high connectivity, as every member is connected to every other member. The *cyclic* network represents sparse connectivity, as each member is only connected to two others (and no two neighbors have any neighbors in common). The impacts of connectivity can be seen with two different comparisons (to be explained shortly). These are helpful to distinguish because in the first comparison, the diversity of the community appears to be a mere necessary consequence of how decreasing connections increases predictability. This in turn highlights how in the second comparison, diversity plays an ineliminable role in the community being more predictable.

For the first comparison, consider any two points that sit on the same horizontal line (have the same $y$-value). I prove in Chapter 1 that a more connected community is expected to use less evidence to achieve the same level of aggregate confidence as a less connected one.

Hence the aggregate opinion increases more quickly for more highly connected networks. On the other hand, notice that the standard deviation is higher for the more connected network. This is also due to the fact that the community is reaching that expected opinion on less information. In this case, the community is less predictable. But the explanation for this concerns only the amount of evidence that is produced. The fact that the less connected community is more diverse seems to be nothing more than a necessary consequence of the connectivity, but does not play a role in why the community is less predictable relative to a particular expected opinion.

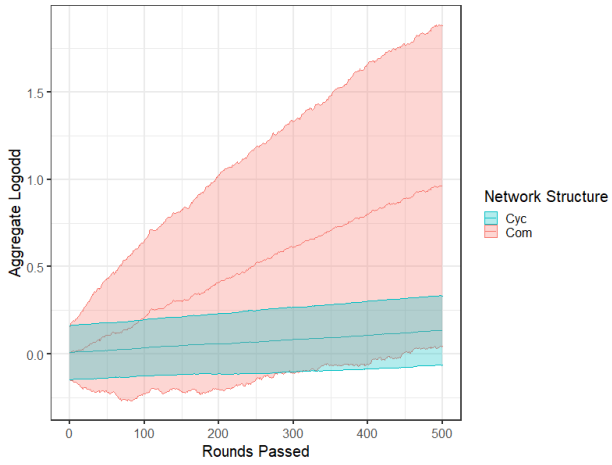



Figure 3.1 The average aggregate logodd over time ($\epsilon = .0005$, $n = 100$)

The other comparison tells a different story. Consider any two points of either network on the same vertical line (with the same $x$-value). Any two such points represent communities with the same amount of evidence produced. But notice that the standard deviation around the complete communities is still larger than the that of the cyclic. So even relative to the exact same amount of evidence produced, the more connected community is less predictable. Why? Because on a complete network, when misleading studies are produced, they are observed by all of the agents, and the aggregate will necessarily reflect this. On a cyclic network, while individuals may have received evidence that pushed them above or below the

average, these do not affect the community as a whole. And as a result, the opinions of the rest of the community balance each other out in a way that means their aggregate is less likely to skew than on a complete network. Hence the more connected and less diverse communities are more susceptible to particularly misleading series of studies, and the individual iterations of the complete communities skew further below and above what is expected. So it is not the case that diversity is merely a necessary consequence, and instead it plays an integral role in the community being more predictable.

But it is not the diversity of opinions alone that is beneficial. The diversity of opinions is beneficial because the agents have a diversity of evidence. Whatever diversity there is between opinions that is stipulated at the outset is quickly washed away by whatever the studies (and their distribution structure) dictate. So that sort of diversity of opinions has little use.[2] If we were to stipulate that agents not move from diverse opinions, the community would be predictable but in a trivial way. So in the case of connectivity, the fact that the diversity is caused by the evidence that the agents hold is crucial aspect. This prompts the following hypothesis:

> **Hypothesis 1:**
> Anything which increases diversity of opinion —via increasing diversity of evidence held— increases predictability.

This hypothesis is more plausible than one which says any diversity of opinions *without qualification* is useful. But it is still false. The next section explains why.

### 3.5 Imprecise Research

In order to see why Hypothesis 1 is false, we need to create a 'bad' form of diversity of evidence. This can be done by making the agents less precise in their data production. One way to generalize the base model from chapter 1 is to give each agent their own *imprecision*

[2]As discussed above, there is a use for that sort of stipulation of opinions in models with preferential research. But as I have already argued, those models merely highlight the value of diversity of research paths, which is not an open question.

*value* which indicates by how much extra noise any given agent's results might distorted. Imprecision like this can occur in the world in a variety of different ways depending on the field. For instance, there might be borderline cases which some researchers are better at distinguishing than others (doctors discerning the cause of death). Some researchers might use better quality or better maintained equipment (old and dirty versus new and clean microscopes). Finally, researchers might make purely innocent calculation errors (which could be as simple tallying results incorrectly) more often than their peers. In each of these cases, the agents' respective levels of imprecision distort their evidence to some degree, but it has nothing to do with their opinions or intentions. It is just as likely to be a distortion in support of their hypothesis as it is against it.[3]

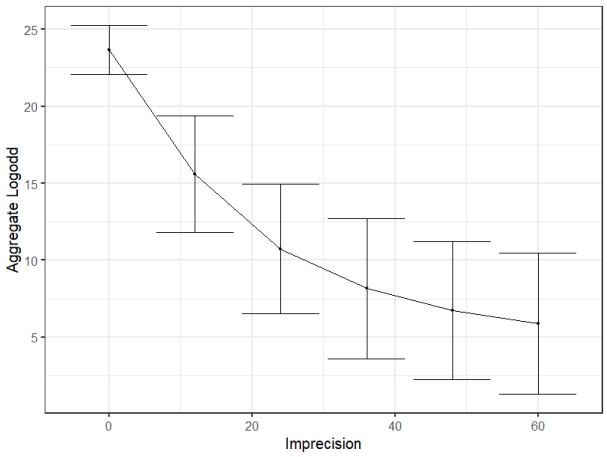
This sort of imprecision can be modelled in the following way. When an agent produces a new sample of data, a number randomly generated in accordance with their level of imprecision, $i$, is added to the number of successes they observe, $k$. More specifically, $k$ is reassigned in the following way:

$$k_{imprecise} \leftarrow k_{original} \oplus randnorm(0, i)$$

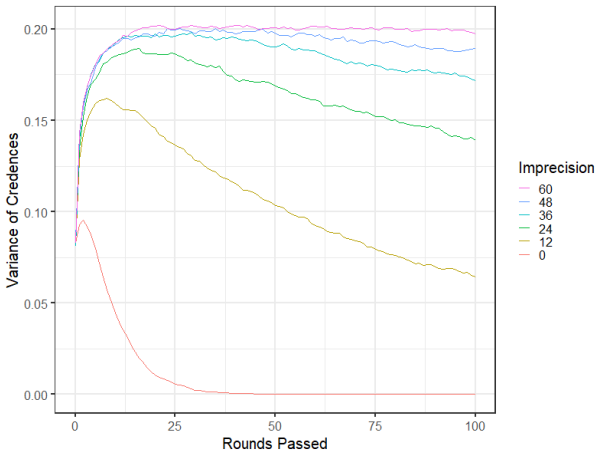
The function $randnorm(m, v)$ yields the nearest integer to a real number generated with mean $m$ and variance $v$. $x \oplus y$ is the sum of $x$ and $y$, but rounded to 0 if it is less than 0 or to $n$ if it is greater than $n$ (the number of trials in the study). Hence $randnorm(0, i)$ generates a number based on a distribution centered around 0 with a variance $i$ — which indicates the level of imprecision for that agent. Notice that an agent's imprecision can cause their results to be distorted negatively or positively (and this does not depend on anything else), and does so in both directions an equal amount. Further, the greater $i$ is (the more imprecise the agent is), the more likely that any given distortion will be larger. Note the results above (as well as the rest of Chapter 1) concerns a perfectly precise community, i.e. for each agent $i = 0$.

[3]This distinguishes imprecision for things like bias, where agents distort evidence in a way that conforms to their view. I consider bias in Chapter 4.

With this generalization of the model, it becomes clear that one can increase the diversity of evidence, and likewise opinions, in a way that has no benefits. Consider Figure 3.2 which shows the results for communities under various levels of imprecision.




(a) The average aggregate logodd and the standard deviation in the $100^{th}$ round




(b) The variance of credences over time

Figure 3.2 Performance as a function of the maximum level of imprecision of the community.

First notice that higher imprecision does in fact increase the diversity of the agents' opinions (Figure 3.2b). This is because any given pair of agents, given higher levels of imprecision, it is more likely that they will observe more disparate results. The chance that one sees results strongly in favor of the new treatment and the other sees results strongly indicating against it is more likely with more imprecision. So this is a case in which a community has more diverse opinions due to a greater diversity of their evidence. Hence, increasing imprecision is a change to the community which satisfies the antecedent of Hypothesis 1: greater diversity of opinion is the result of greater diversity of evidence.

However notice that the community is less predictable when there is more imprecision (Figure 3.2a). That is because any given run of the simulation is more likely to observe a series of particularly high or low frequencies and thus be further from the mean of all

simulation runs for their network. This is because the studies themselves are more likely to have such high or low success rates. So it is more likely that a particularly misleading series of studies occurs for an imprecise community than a perfectly precise one.

This shows that Hypothesis 1 is false. It is not the case that any way of increasing the diversity of opinions by increasing diversity of evidence held will be beneficial. One can increase the diversity of evidence held simply by increasing the diversity of evidence produced. But increasing the diversity of the evidence produced makes the issue at hand more challenging, and there is no upside to doing so (at least that one can see here). This leads to the following, more plausible hypothesis to consider:

> **Hypothesis 2:**
>
> Anything which increases diversity of opinion —via increasing diversity of evidence held *without increasing the diversity of evidence produced*— increases predictability.

This hypothesis differs from the first solely by inclusion of the qualification that the diversity of the evidence produced is not increased. This is meant to guard against counterexamples like imprecision. One might think that changes made that allow the community more diverse pools of evidence, without somehow forcing the evidence produced to be more disparate, could only be beneficial to the community by allowing them to capture the diversity that is already there in the evidence. That is what happens in the case of increasing connectivity. The diversity between agents' pools of evidence increased, but it does so without increasing the diversity of the evidence itself. One might likewise think that any changes to the community itself (not its research) which cause increased diversity will also cause increased predictability. Changes to the network structure offer perfect examples of this, because they do not affect the individuals' scientific research, just with whom the data is shared. However, the next section shows that this line of reasoning, and likewise Hypothesis 2, is wrong.

### 3.6 Network Centrality

Like Hypothesis 1, Hypothesis 2 is too strong of a claim. One must be more careful in how they diversify opinions for that diversity to be useful. Simply guarding against increasing the diversity of the evidence produced is not enough. This can be seen by analyzing another network change that is distinct from network connectivity discussed above.

Recall that the connectivity of the network indicates the amount of connections on that network, but there are many more ways to characterize networks. Likewise, there are many more possible changes to networks than simply adding or subtracting connections. In particular, one can consider the way that those connections are arranged. This section concerns what happens when networks are more or less evenly distributed, what I call the *centrality* of the network. This can be explained by comparing two networks. Recall that on a cyclic network every agent has 2 neighbors apiece, and no two neighbors share a third in common. On a *star network*, one agent is connected to every other agent, and there are no other connections. Note the star network has nearly the same amount of connections as the cyclic (just 1 fewer, for any population size above 2). However the star network's connections are more centralized. That is because one agent has a maximal amount of connections while the rest only have 1 apiece.

We can expand on this by introducing two parameters for networks. The number of *stars* indicates the number of agents on a network that are connected to everyone. (So the star network is defined by having a single star agent and no other connections.) The rest of the agents are called *perimeter agents*. The second parameter is the network's number of *cycles*, which indicates the number of connections that each perimeter agent has. A network having $k$ cycles means that each of its agents is connected to the nearest $k$ agents on either side of them (when situated on a ring, with the star members placed in the center). A series of networks that differ by their stars and cycles is shown in Figure 3.3. Notice that on every row, the networks have about the same number of connections. As one moves from left to right on any given row, the level of centrality increases. With these networks in mind, we

can decouple the effects of network centrality from those of connectivity.

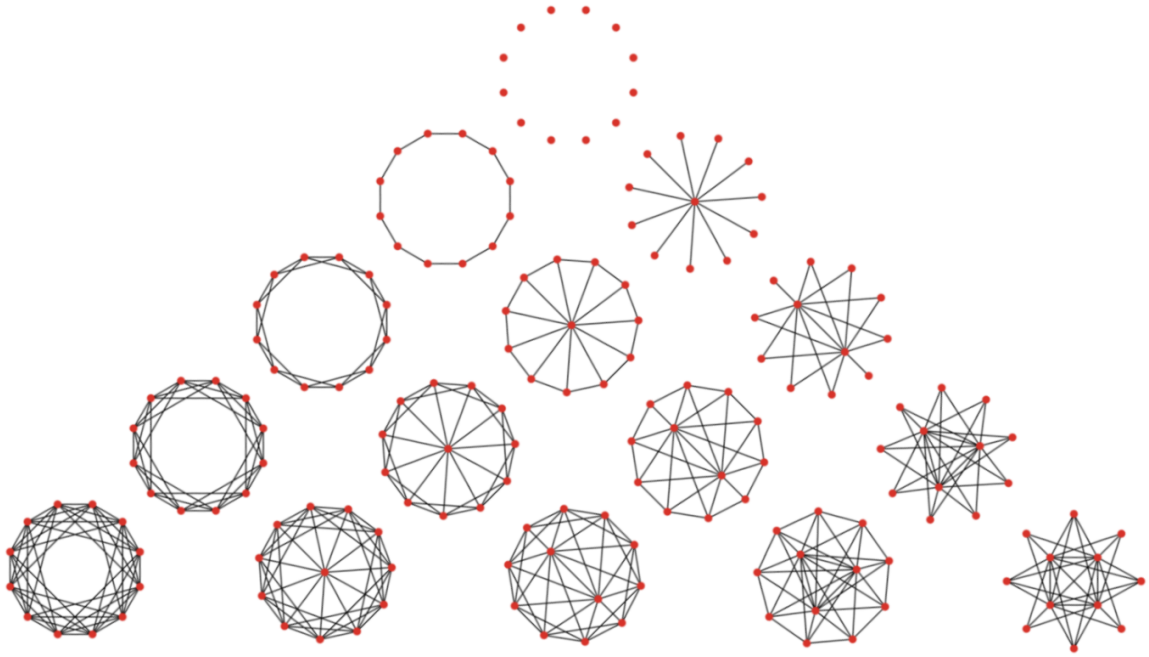


Figure 3.3 Networks that differ by amount of cycles and stars. For any network, moving down and to the left adds a cycle. Moving down and to the right adds a star. Each adds roughly the same amount of connections.

Figure 3.4 shows the results for all of the networks on Figure 3.3 after 100 rounds of the simulation. First, notice that the centrality of connections does not affect the speed of network. The speed is only affected by the amount of connections, because (since agents are all perfectly precise) the aggregate opinion depends solely on the amount of information that is being produced and shared. Since centrality does not affect either of these, it has no impact on the speed at which the aggregate increases. Notice that the only difference between speed performance for networks on the same row (from Figure 3.3) can be attributed to the fact that they differ by 1 or 2 connections, indicating that the *amount* of connections is the determining factor. This is useful for distinguishing between the effects of connectivity and centrality, but it does not say much about whether diversity (of any form) has value.

For that, we must turn to the two variance measures (predictability of the aggregate and diversity of opinions).

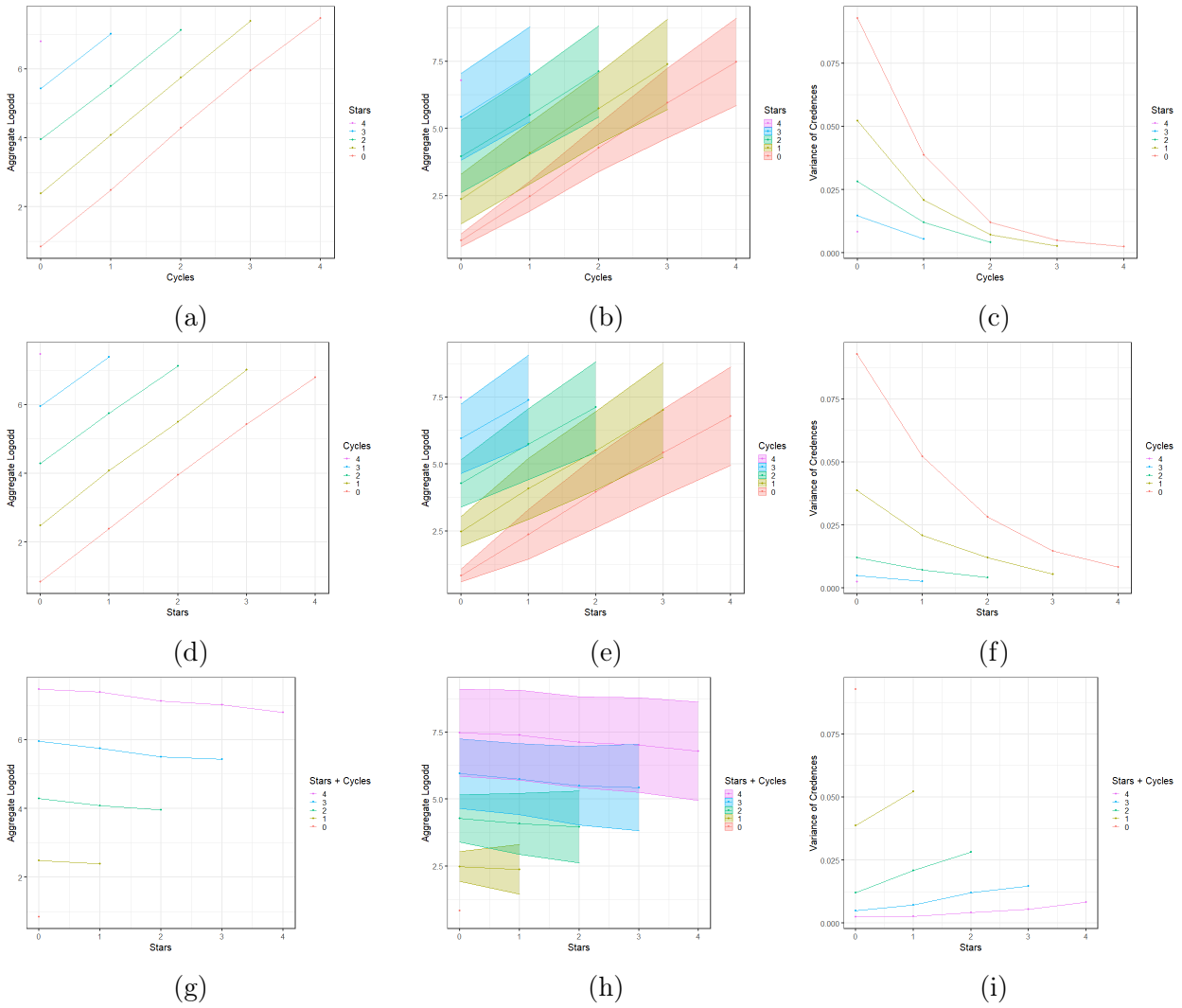



Figure 3.4 Performance for the 15 networks from Figure 3.3. The top row displays results when holding constant the number of stars and varying cycles; the middle row holds constant cycles and varies stars; the bottom row holds (roughly) constant the number of stars + cycles (i.e. roughly the total number of connections) and trades-off between the two. The left column represents the average aggregate credence; the center column is the same with the standard deviations imposed; the right column is the variance of the credences.

The impacts of network centrality on the unpredictability and the diversity of the com-

munity are more enlightening. Recall that increasing diversity of opinions *by decreasing connections* leads to a more predictable community. In this way, the results of Chapter 1 can be seen as imposing a trade-off between the variance within the model at any given time (the variance of the individuals' opinions) and the variance between (the aggregates of) the runs of a simulation. More variance within the community means less variance between runs of communities. More variance between runs of communities means less variance between the individual members on any given run. One might think that this is a general trade-off, which in this case is being revealed by the amount of connectivity, but which holds based on any change to the network itself.

However this is in fact false. Not all causes of variance between individual opinions curtails the variance between runs of the simulation. This can be seen with how the centrality affects both forms of variance. It turns out, increasing centrality both increases diversity *and* unpredictability. To see why, consider both aspects in turn, starting with the diversity of opinions within a community.

It is counter-intuitive that increasing centrality actually makes a community more diverse. Stars on a network make it so everyone in the community shares at least some overlap in their evidence. Consider a cyclic versus a star network. In the cyclic, each agent has some overlap of evidence with each of their neighbors, but there is no evidence which is seen by everyone. On a star network, half of everyone's evidence is identical. One might think that this implies the more centralized network must be more unanimous. But this is in fact false. Increasing centralizing decreases the level of unanimity.

Why? Consider the effect of any given connection on a network. A connection on a network makes it so that both connected agents are closer to the community average.[4] That is because their connection makes it so each hears more evidence present in the community, and so as a result their opinion is closer to the aggregate opinion because it reflects more of

[4]There are special cases of extremetization explored in Chapter 3, where in particular rounds, agents being connected *increases* overall variance. But the key here is that *on average*, a connection brings agents closer to the average of the community.

the same evidence. On a cyclic network, each agent has two connections. On a star network, everyone except the star has 1 connection. That means each of perimeter agents on the star network will be further from the aggregate than the agents on the cyclic network will be — simply because they each have less connections.

One might wonder why the high amount of the star connections does not balance out the effects of the lack of connections for the perimeter agents. That is, there is (roughly) the same amount of connections, so why is there not the same collective decrease in the amount of variance in the community? The reason for this is because additional connections have *diminishing returns* with regard to the amount of diversity they curtail. That is, every additional connection that is added to an agent has less of an effect on bringing their credence closer to the aggregate. The upshot of this is that if one wants to maximize the amount of cohesion on a network, they ought to distribute connections as evenly as possible. Doing so is the best way to optimize the diversity-limiting power of each connection. For any network with unevenly distributed connections, it can be made less diverse by playing Robin Hood: take connections from those that have them, and give to those that do not.

On the other hand, higher centrality increases the unpredictability of the network. That is because increasing centrality makes it so the community's aggregate opinion reflects more closely a smaller subset of the produced research (those produced by stars). This makes the aggregate opinion more susceptible to the possibility that one or more stars gets a particularly skewed series of results throughout their iteration which skews the aggregate. Decentralizing connections protects the community against over-relying on any agent's evidence.

In other words, while no agent ever double counts a study on a network, we can think of connections as making it so the community as a whole counts multiple studies multiple times. We can see this in the complete and cyclic networks above. For the complete network, the community's aggregate opinion is determined by updating on each study N many times over (because every one of the N agents updates on the study, and each impacts the aggregate). For the cyclic network, each study is counted three times (by the agent that produces it and

the two neighbors). But in both of those cases, the amount of overcounting is the same for every study, because the connections are evenly distributed. Therefore, no agents' evidence is overcounted any more by the community as a whole, and so the aggregate opinion of the community evenly reflects everyone's evidence. On networks with stars however, this is not the case. The stars' studies are counted by the community more than any of the perimeter agents' studies are. This makes it so the aggregate opinion of the networks with starts are more distorted to reflect their evidence than any other agents'.

This shows that increasing diversity of evidence held —even without increasing the diversity of the evidence produced— does not guarantee a more predictable (or faster) community. In this case it leads to a less predictable one. This shows that the trade-off between variance within communities and between them is not general to all network changes. One can change a network to make it so the members of the community are more diverse but less predictable.

Since centralizing network connections increases the diversity of opinions, it satisfies the antecedent of Hypothesis 2. But because it increases the unpredictability, it shows that Hypothesis 2 is false.

Just like the failure of Hypothesis 1 motivated Hypothesis 2 (i.e. protecting against causing extra variance in the evidence itself), the failure of Hypothesis 2 helps motivate a third hypothesis. The discussion of centrality shows that what is truly at issue is how the community treats each of its sources of evidence — how many times each study is counted. It shows that what is important is that the community members' distribution of opinions accurately reflects the distribution of the evidence provided (and does not give disproportionate weight to any set of studies). In doing so, the distribution of the community's opinions more accurately reflects the distribution of the evidence itself. That is, for the networks with stars, there is less variance in the opinions of the community than there is between the studies that are produced. That makes it so the community's aggregate opinion is in a sense *blind* to the true variance of the studies. Alternatively, when the diversity is caused by decreasing connections, this is done in a way that reveals the diversity of the studies. That is, when di-

versity is caused by decreasing connections, it makes it so the aggregate community opinion *sees* the diversity of the evidence. This prompts the following hypothesis:

> **Hypothesis 3:**
>
> Anything which increases diversity of opinion —*by reflecting* the diversity of evidence produced— increases predictability.

A community's distribution of opinions *reflects* the distribution of its evidence to the extent that it has the same variance (and corresponding average). So for a community's distribution of individual opinions to perfectly reflect the studies, it must have the same variance. A community is worse at reflecting the variance of studies the greater disparity there is between the variance of opinions and the variance of the studies produced.

This hypothesis is importantly different from Hypothesis 2 (and it is important to distinguish them since this one is likely true). Hypothesis 2 rules out changing the distribution of the evidence itself. For Hypothesis 3, this is strengthened to make it so that not only should that distribution not be altered, but it needs to be reflected. And as the above shows, there is a crucial difference between not altering the distribution and better reflecting it.

This hypothesis gets at *why* diversity caused by network connectivity has the usefulness it does. Only by doing so can the aggregate opinion of the community *see* the diversity of the studies. But further, it shows that the underlying insight is in fact a very intuitive claim centering around what happens when the diversity of studies produced is ignored. This is already accepted on an individual level, that one must appreciate there is variance to the studies she observes (otherwise, there would be no point to more than one sample). However this result demonstrates the same recognition is useful at a societal level.

In the next section, I provide a final model complication which helps support Hypothesis 3.

### 3.7 Intransigence

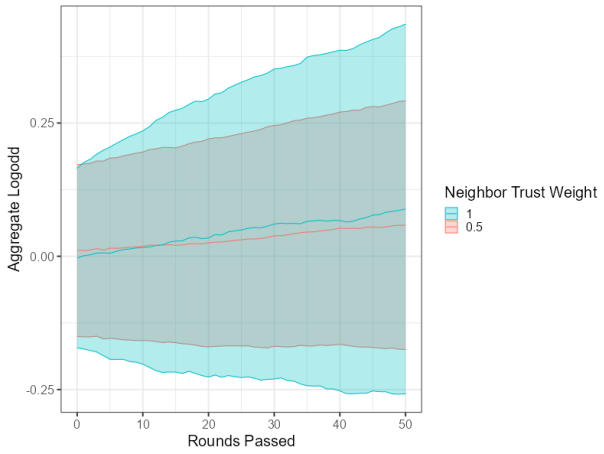
To help support Hypothesis 3, we can look at a simple mechanism which causes a diversity of opinions by allowing them to better reflect the diversity of evidence produced. Consider a community consisting of members that are slightly skeptical of one another. While agents take their own studies at normal value, they discount those they hear of their peers by 50%. (Note, the same can be demonstrated with any level of discount.) These sorts of strategies are more relevant for more complicated scenarios like where imprecision and bias are in play (which I consider in Chapters 3 and 4 respectively[5]), but it is worthwhile to consider under simpler conditions first because, as can be seen here, it has utility even without the normal motivations. (So as with the rest of the results outside of Section 4, agents here are perfectly precise.)

Notice that when agents trust their own evidence more than those in their community, their opinions in turn reflect their evidence more than the others'. Hence, as seen in Figure 3.5b, discounting neighbors evidence in this way creates a diversity of opinions. Further, it does so in a way that reflects the diversity of evidence produced. As the amount of trust increases, the amount of correlations between agents increases, and likewise the variance between their opinions decreases. Therefore, imposing this trust strategy constitutes a change which satisfies the antecedent of hypothesis #3: it increases variance by reflecting the diversity of the evidence produced.

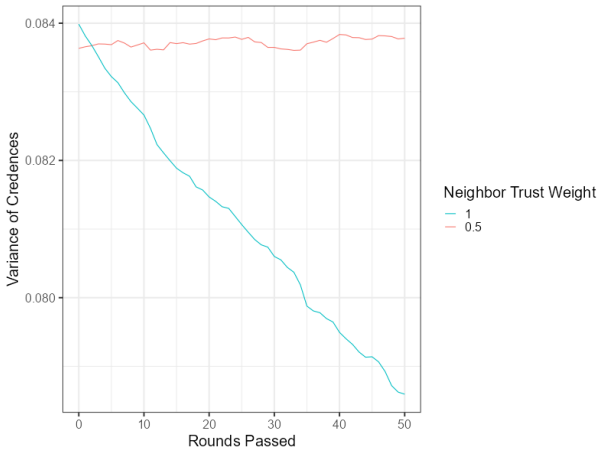
It also satisfies the consequent, as seen in 3.5a. The higher amount of diversity makes it so the community's aggregate opinion is more predictable. This is because the community's aggregate opinion is reflecting the diversity of the evidence produced better than with full trust. When agents do not distrust each other's evidence, it is more likely that the community is more susceptible to misleading studies which skew the entire average.

This does not conclusively prove Hypothesis #3 is correct, but it provides strong sup-

[5]In those chapters, the version of skepticism I consider is full skepticism for reasons made clear there. However that strategy is equivalent to removing all connections, which I have already displayed results for. It is more helpful to consider a fundamentally different sort of mechanic that has the same effect.




(a) The average aggregate logodd and the standard deviation over time




(b) The variance of credences over time

Figure 3.5 Performance for fully trusting neighbors versus discounting their evidence by 50% (complete network)

port for it. This change (affecting the amount of trust) and the one from Chapter 1 are structurally very different — one affects the network itself while the other concerns the updating of information alone. Still, they both can be seen to demonstrate the mechanism that Hypothesis #3 focuses on.

## 3.8 Conclusion

The above highlights that the utility of diversifying opinions concerns the way in which it allows the community as a whole to appreciate the diversity of evidence produced. That is, when the diversity of the evidence is reflected by the diversity of the society, as opposed to every member having an essentially duplicate scientific process, the community becomes more predictable. This demonstrates a different sort of division of cognitive labor than is usually discussed. In this sort, labor is understood as properly handling various pools of the community's evidence, not just different research paths. So we can think of the distribution of cognitive labor as not merely applying to scientists with levels or areas of expertise, strategies or projects, but also in terms of how evidence itself is distributed.

At this point, it is important to note that this is a community level concern, one which

concerns the expectation of a community's aggregate opinion. Each individual, regardless of the network structure, has the same expected opinion *and the same likelihood of achieving it.* As I prove in Chapter 1 (assuming perfect precision), each agent's individual opinion is determined solely by the amount of evidence they acquire. An agent's opinion is not dependent on whether another has the same evidence that they do. What changes is the predictability of the *aggregate opinion.*[6] The concerns here demonstrate challenges for achieving particular desiderata for the community's opinion, but they are —in this limited model— irrelevant to the performance of the individuals themselves. Therefore this demonstrates an interesting angle on the independence thesis introduced by Mayo-Wilson et al., 2011. The thesis states that achieving certain desiderata for members of a group is independent from achieving desiderata for the group. Considering these two network changes on this model demonstrates how this is the case. The trade-offs above concern achievement of group desiderata but not desiderata of its members.

Does this undermine the importance of the above results? One might think that the aggregate community opinion is a pretty trivial notion if it does not ultimately mean something to the individuals. That is, if dynamics that affect the aggregate for whatever reason do not affect individuals, then why should we (or the individuals) care?

There are two reasons. First, many philosophers take the notion of a group mind seriously. On this view, the group has its own cognitive and likewise epistemic states. The aggregate opinion discussed here then, is not merely a metric of the individuals, but a metric of an aggregate entity. With this understanding, the group itself has its own epistemic desiderata, and the predictability of its opinion is likely one of them (just as an individual's predictability is a desideratum). Hence it is worthwhile to consider how a group's epistemic desiderata are achieved for their own sake. It is in this way that the results can be understood as a defense of the independence thesis.

Second, and less tendentious, is that even if someone is not convinced that a group can

[6]Increased connectivity does increase the speed at which agents acquire evidence, but it does not disproportionately affect the predictability of their individual opinions as it does to the aggregate opinion.

have its own bonafide mind, one still must make some room to appreciate the importance of aggregate opinions. It is important to remember that the model at hand is a *minimal* one, where as many complications as possible are removed (and only added back one by one). This is done to discover the underlying mechanisms involved, but the upshot is that these results hold for much more complicated models as well. In Zollman's model for instance, the above analysis explains why the community as a whole is more likely to abandon the optimal research path. This sort of result is clearly impactful for the agents themselves. And so is any mechanism introduced which involves the community making some decision based on the aggregation of its members opinions (e.g. any form of democracy). The results are the same if we merely takes a binary vote as opposed to aggregating opinions. The point is that scientific communities, juries, and board rooms all have members who are impacted by the aggregate epistemic state of their respective communities. So the results here do apply specifically to only the aggregate, but it is painfully easy to include additions (with real world precedent) that make such aggregate opinions (and their predictability) bear on the individuals as well. Zollman's model with preferential research is an example of this.

The upshot of all this is that a community can benefit from its members having more diverse opinions if the diversity of those opinions is a reflection of the diversity of evidence produced.[7] This shows that the diversity of opinions in the community does impact a community's predictability. But further, it identifies the mechanism of why this diversity is beneficial. In short: Reflecting the diversity of evidence in the diversity of community opinions increases the predictability of a community.

[7]With regard to this specific model, this statement can be strengthened to an *if and only if*.

# 4. SILOING & EXTREMETIZATION: INVESTIGATING THE INFLUENCE OF EVIDENCE CORRELATION AND TRUST ON SCIENTISTS' OPINIONS

## 4.1 Overview

This chapter explores how variance in scientific evidence can lead to variance in scientists' opinions in two distinct ways: *siloing* and *extremitization*. Siloing occurs when agents generate evidence from identical distributions but temporarily diverge in their opinions due to independent pools of evidence. Extremitization occurs when pools of evidence are correlated, causing agents to pull each other further from the average than they otherwise would go. This chapter investigates *trust strategies* as ways of mitigating or exacerbating this diversity, while also considering how predictable these strategies make the community. The results show that trusting those one disagrees with and distrusting those one agrees with (*heterophily*) can limit the variance caused by scientific evidence while keeping the community more predictable than alternative trust strategies. This study contributes to our understanding of the value of epistemic diversity and the potential costs and benefits of various trust strategies in scientific research.

## 4.2 Introduction

I show how variance in scientific evidence creates variance in scientists' opinions in two related, but crucially distinct ways. First, even when agents generate evidence from identical distributions, the agents can temporarily diverge in their opinions if those distributions are independent (*siloing*). Second, when the pools of evidence that agents update on are correlated (e.g. they share evidence with each other, and so both form their opinions on largely the same evidence), either one producing outlier studies can pull the other further from the average than they otherwise would go (*extremitization*). I show that these phenomena can operate without any sort of bias or malfeasance and result purely from the variance of studies alone. Appreciating these two mechanisms — and how they are different — motivates different strategies of trust as ways of mitigating or exacerbating this diversity. In particular, I consider five different trust strategies and show how each one handles the variance produced by scientists. I show a trust strategy that has so far not been considered in the literature, *heterophily*, is able to mitigate both forms of diversity by creating correlation between agents that disagree and creating independence between those who agree.

This project builds off Chapters 1 and 2 where I show that inducing diversity of opinions can cause a community to be more predictable. This contributes to our understanding of the potential value of epistemic diversity by uncovering the dynamics of its causes and effects. In those chapters I show that diversifying opinions (in the way I describe) is sufficient for increasing predictability, but one might wonder if it is necessary. That is, for a community to be predictable, must they diversify the opinions? In this chapter, I show how a community can achieve the same sort of predictability *without* allowing for diversity of opinions. I consider trust strategies as ways in which agents can navigate the variance of each other's data to achieve their various epistemic ends.

I am interested in strategies for agents calibrating their levels of trust for one another, what I call *trust strategies*. These have also been referred to as *norms of trust*, but I take the term 'strategy' to be more neutral. Both strategies and norms are rules by which agents

conduct their action. However, a *norm* suggests a *norm*ative status — that the rule in question is a *good* rule or the *right* one. I refrain from these sorts of commitments and take this project to be *descriptive*. That is, I unfold the dynamics of various trust strategies to show their potential causes and effects. In doing so, I reveal how strategies can be useful or beneficial *conditional* on certain plausible, yet dischargeable assumptions about epistemic desiderata. The point is not to defend a normative claim specifically, but to defend descriptive claims which have significant bearing on normative considerations. Theoretical economics, particularly game theory, functions in the same way.

This project contributes to others in the literature which use computer models to test various strategies (or what they refer to as norms) of trust. Mayo-Wilson, 2014 and K. J. S. Zollman, 2015 are two projects that both test various trust strategies in the context of certain forms of unreliability (like potential miscommunication or mishandled evidence). They both consider the aggregate performances of the community, which allows them to make important insights about what can be expected on average in the community. However they leave open considerations of both diversity and the predictability of the community. That is, while those projects discuss which value the communities are on average expected to reach, they do not discuss how far away from that value the communities might get. They are also silent on how spread out the community's members are in regards to their individual opinions. In addition to considering more metrics, I also consider a trust strategy that has so far been ignored. In this strategy, agents trust others to the extent that they *disagree* with them. I show that this strategy performs better with regard to the new metrics I introduce (predictability and diversity).

In the next section, I discuss briefly the other two projects in the field. This is followed by an explanation of my model and why it is suitable for this investigation. The results are broken into two sections: First, I analyze the behavior of the strategies on a network with sparse connectivity, which presents a nice environment to explain both siloing and extremitization. Second, I consider how these dynamics are affected by the addition of more

connections. I conclude by explaining what has been shown about how variance propagates, and how trust strategies interact with it.

## 4.3 Background

### 4.3.1 Mayo-Wilson

Investigating trust strategies using computer models begins with Mayo-Wilson, 2014. In that project, Mayo-Wilson models a community faced with multiple questions, and each agent only researches one question, called their *expertise*. Inaccurate opinions are caused by randomness in the evidence that is generated as well as the potential for miscommunication. Each agent is just as reliable with regard to their expertise, and there is the same chance of miscommunication on each message transmission.

Mayo-Wilson shows that only when experts are well enough distributed and their testimony trusted enough is the community able to achieve its epistemic ends. This is an interesting result for how we should distribute sources of evidence (and likewise evidence itself). It shows the importance for the community of having its evidence diversely distributed. When most of the evidence for a particular question is restricted to a relatively small portion of the community, the community as a whole performs worse. Experts on a question benefit less from being connected to other experts than non-experts do.

While this result is interesting, the model leaves open what happens when agents in a community have different levels of unreliability. On Mayo-Wilson's model, there is no reason to differentiate agents except based on their expertise. One might accept that trusting experts is a good idea, but still wonder what they should do when they suspect a particular scientist is less reliable than others.

### 4.3.2 Zollman

Zollman's model in K. J. S. Zollman, 2015 gives us an insight into how trust strategies perform when agents have different levels of reliability. On his model, agents also are faced with a number of different research questions. However, contrary to Mayo-Wilson's, agents

do not have an expertise. Instead a random portion of the community receives a signal from nature concerning one of the questions. Each agent, depending on their respective level of reliability, correctly or incorrectly reports the answer (TRUE or FALSE) to the question they hear about with others in their community.

Since Zollman's model does not include expertise, the strategies do not concern expertise either. Instead, he is concerned with four strategies (versions of which I also consider for my model):

***Credulism***:

Agents trust each other fully.

***Homophily***:

Agents trust each other in direct proportion with the extent that they agree.

***Imprecision-avoidance***:

Agents trust each other less depending on their likelihood of sharing misleading information.

***Skepticism***:

Agents do not trust each other at all.

Zollman refers to homophily and imprecision-avoidance as *subjective reductionism* and *objective reductionism* respectively. This is in reference to the idea that agents ought to calibrate their levels of trust to the appropriate level depending on the level of reliability of the source of information. Objective reductionism is described by Zollman as a "crystal ball" type of strategy, because it presupposes that agents have perfect insight into the level of imprecision that each other has. It seems unlikely that an agent knows perfectly their own level of imprecision, let alone that of others. But the strategy is meant to be instructive by showing what is theoretically possible. Subjective reductionism, on the other hand, only relies on what agents' opinions are. It is more plausible that this is publicly available, and so this is meant to represent a more realistic strategy for agents to calibrate their trust based on information that they have access to. Hence this is a subjective way for agents to reduce

their neighbors' reliability to a particular metric.

I avoid the term "reductionism" because there are a lot of ways to calibrate levels of trust, and it is not clear that either of the ones above are the correct version. For subjective reductionism especially, it is helpful to recognize that this strategy is a version of homophily, which is a widely discussed tendency of agents to associate with those of like minds. There are a lot of other ways in which someone might try to calibrate their trust of others, whether that be based on each other's opinions (one of which I introduce below) or other information available to the individual. But even with regard to objective reductionism, it should not be assumed that Zollman's version (distrusting the imprecise) is an ideal strategy in the sense that it is perfect or the best. It is ideal in the sense that it depends on assuming agents have perfect information with regard to each other's level of imprecision. But it is not a given what one should do with that information. (In fact, I show other strategies outperform this one with regard to certain metrics.) There are plausibly other objective reduction strategies that are based on different information. For an example, a strategy might also consider how many connections an agent has. I do not test such a strategy here,[1] but that it could exist shows that "objective reductionism" is a broad camp of strategies, not a single one. I refer to his version of objective reductionism as imprecision-avoidance, because it is more informative: agents avoid imprecision.

Zollman shows that in his model, credulism is on average able to achieve more true beliefs than any other strategy, even imprecision-avoidance. The latter strategy does have a higher portion of true versus false beliefs by a significant margin, but credulism has a higher ratio than both homophily and skepticism. Zollman uses these results to argue that credulism in such scenarios seems to be the best strategy to employ. With it an agent is expected to have the highest amount of true beliefs, and out of at least the plausible strategies (i.e. not imprecision-avoidance) the fewest amount of false beliefs.

[1]The reason is because such a strategy involves a number of complications the dynamics of which need to first be established independently. In future research, I plan to investigate strategies like this one, but this project is a necessary first step.

Zollman's model provides a strong starting point for this project. It introduces a sense of intrinsic unreliability where each agent has a chance, unique to them, of distorting their evidence in a particular way. The strategies are the right sort of strategies we should begin by testing. Credulism and skepticism give us two baselines at either extreme. Imprecision-avoidance focuses on what appears to be the source of the problem. And homophily is concerned with distances between opinions, i.e. diversity, which is one of the central focuses of this paper. By assessing the potential outcomes of employing these strategies, we can better understand the benefits and limitations of each approach.

Zollman's project also presents multiple ways in which research can be furthered. For instance, there are more strategies to consider and metrics to judge them with. We can also focus more closely on a single question for which agents need to collect evidence over time. This allows us to see the dynamics of trust strategies with regard to data concerning a specific question, which presents its own sources of diversity. I formalize this in the next section.

## 4.4 Model

My model concerns a network of agents that share statistically generated evidence to update their opinions concerning a particular problem. The value of this model is that it applies to the broad range of scientific communities that rely on random sampling, and it depends on minimal epistemic assumptions. The model also allows us to consider the ways in which evidence itself can be diverse. This turns out to be integral in how diversity in a community is propagated, and in turn presents a new way for trust strategies to be evaluated. I will first explain the basics of the model before generalizing it to account for imprecision and then trust. Subsequently, I explain the metrics used to explore the model.

It is important to remember than none of the specific values in the model are terribly important, as they can be tuned at will. The interesting features come from the *dynamics* of the model. These are general features of the epistemic situation and can be explained without reference to any particular values. The dynamics explained in this chapter (just like

Chapters 1 and 2) hold for all parameter settings.[2]

### 4.4.1 Setup

The problem in question can be thought of as whether or not a particular frequency in nature is above or below a given threshold. One can think of this as doctors determining if a medical treatment's success rate is higher than a known standard, but (as I discuss in Chapter 1), there are many interpretations of the model. All that is important is that the agents used statistically generated evidence to update their opinions. The community consists of $N$ many agents that each produce $n$ new data points every round. The observed frequency of a study is $\frac{k}{n}$ where $k$ denotes the number of successes for the study. The $\frac{k}{n}$ of any given study is likely above or below the true frequency, but overtime the aggregate frequency of the studies reflects it. Agents research every round, and so this entails the amount of evidence that is produced each round is constant. This ensures that the dynamics that unfold are due purely to the flow of information, and not how or if it is generated.

Every round $N \times n$ many data points are produced in the community, but the amount of those data points that any given member of the community observes is dependent on the network structure. Agents only observe the studies of those they are connected to on the network. Chapter 1 concerns how the amount of connections on a network affects its performance, and below I explain how those results have bearing here.

Imprecision can be modelled by giving each agent their own *imprecision value* which indicates by how much extra noise any given agent's results might be distorted. Imprecision like this can occur in the world in a variety of different ways depending on the field. For instance, there might be borderline cases which some researchers are better at distinguishing than others (doctors discerning the cause of death). Some researchers might use better

[2]This project is interested in the dynamics which should never be identified with any particular metrics. Sometimes finding or calibrating the right metric to observe the dynamic for a certain simulation might be difficult. But there is no *a priori* reason why we should suspect any given metric we choose has an insight into a model. If none of our telescopes can see Neptune, but we know it must be there, we simply need better telescopes. We do not just *assume* that our telescopes will show us everything of importance. Dynamics first, metrics second.

quality or better maintained equipment (old and dirty versus new and clean microscopes). Finally, researchers might make purely innocent calculation errors (which could be as simple tallying results incorrectly) more often than their peers. In each of these cases, the agent's level of imprecision distorts their evidence to some degree, but it has nothing to do with their opinions or intentions. It is just as likely to be a distortion in support of their hypothesis as it is against it.[3]

This sort of imprecision can be modelled in the following way. When an agent produces a new sample of data, a number randomly generated in accordance with their level of imprecision, $i$, is added to the number of successes they observe, $k$. $i_j \leq I \leq 1$, where $I$ is the maximum level of imprecision for members of the community. More specifically, $k$ is reassigned in the following way:

$$k_{imprecise} \leftarrow k_{original} \oplus randnorm(0, i) * n$$

The function $randnorm(m, v)$ yields the nearest integer to a real number generated with mean $m$ and variance $v$. $x \oplus y$ is the sum of $x$ and $y$, but rounded to 0 if it is less than 0 or to $n$ if it is greater than $n$ (the number of trials in the study). Hence $randnorm(0, i)$ generates a number based on a distribution centered around 0 with a variance $i$ — which indicates the level of imprecision for that agent. This value (which is almost always less than 1, because $I$ is less than 1) is multiplied by the size of the samples taken in order to make the distortion of an agents' imprecision relative to the size of the tests taken.[4] Notice that an agent's imprecision can cause their results to be distorted negatively or positively (and this does not depend on anything else), but does so in both directions an equal amount. Further,

[3]This distinguishes imprecision for things like bias, where agents distort evidence in a way that conforms to their view. I consider bias in Chapter 4.

[4]One might wonder whether this is appropriate, since larger studies might be thought to wash out the effects of an agent's imprecision. Hence one might think that the effects of imprecision should not scale with the size of the tests. But this is wrong. If an agent is bad a determining borderline cases, uses poor equipment, or makes calculation errors — all for individual data points —, then the potential for these same failures would be present for every particular patient or data point considered. Yes, the agent should take more studies, and doing so will eventually wash out the effects of his imprecision, but his imprecision is still a danger for every data point, and likewise the effects of imprecision scale with the size of the tests.

the greater $i$ is (the more imprecise the agent is), the more likely that any given distortion will be larger. Note the results above (as well as the rest of chapter 1) concern a perfectly precise community, i.e. for each agent $i = 0$.

Trust can be modelled by having agents discount the evidence they hear from their connections. Each agent $j$ assigns a weight between 0 and 1 for each connection $l$ that they have, denoted $w_{jl}$. Before updating on the evidence they hear, an agent multiplies both $k$ and $n$ by the $w$ they have for that agent. Hence, the level of trust does not change the frequency that the study indicates ($\frac{k}{n} = \frac{w*k}{w*n}$), but it does decrease the level of impact of that study on the agent's opinion. (Note a $k = 60, n = 100$ study impacts an agent's opinion more than a $k = 6, n = 10$ one.) This allows us to redefine the above strategies in terms of $w_{jl}$.

***Credulism:***

$w_{jl} = 1$

***Homophily:***

$w_{jl} = 1 - |c_j - c_l|$

***Imprecision-avoidance:***

$w_{jl} = 1 - |i_l - I|$

***Skepticism:***

$w_{jl} = 0$

***Heterophily:***

$w_{jl} = |c_j - c_l|$

Note the fifth strategy is not one that Zollman (or anyone else) considers. But I will show below that it is more interesting than the rest. This strategy is similar to homophily in that agents determine their level of trust based on the distance between their credence and the other agent's. The difference is that it works in essentially the opposite way: agents are trusted less the closer their credences are.

It is important to note that while these strategies are articulated in terms of credence, parallel strategies can be formulated with other representations of the agents' epistemic state, including logodds (which are a logarithmic transformation of credences) as well as simple estimates (which require less epistemological machinery). In future work I plan to study the intricacies of using one representation versus another, but for now, all that is important is that the dynamics discovered here hold for more than just the metrics that I demonstrate here. The explanations for these dynamics will make this clear, as they rely purely on whether or not there is correlation with the agents' evidence, and not what they do with that evidence (so long as there is some consistent story about what each agent does).

### 4.4.2 Metrics

As mentioned above, both Mayo-Wilson and Zollman use metrics that consider the expected performance of communities, as understood as their average performance over many iterations. Metrics like this, which ignore variance, are indispensable for analyzing a community and provide insight into what individuals can expect. However it is also important to consider different metrics that rely on *variance*, as they characterize the communities in ways that are hidden with regard to averaging alone.

In particular, we can gauge the variance between the aggregate opinions of different runs of the simulation. Doing so indicates how *predictable* the community is. This is important for any community that might depend on community level decisions, or when complications to the model are introduced (like preferential research — where agents only research when they find it worthwhile) that make it so agents depend on each other for evidence to be generated. The variance between runs allows us to capture how spread out particular runs are as opposed to merely considering the average performance of the runs.

In addition, we can gauge the *diversity* of the agents' opinions. I have already mentioned how the statistical nature of the evidence allows us to consider a variance in evidence, but it is just as worthwhile to consider a diversity in the opinions. In chapter 2, I show how certain ways of inducing a diversity of opinions can be useful by increasing predictability.

Here however, I investigate ways in which a community might strive for predictability while also limiting their diversity (i.e. without sacrificing some agents to be less accurate than others).

## 4.5 Results

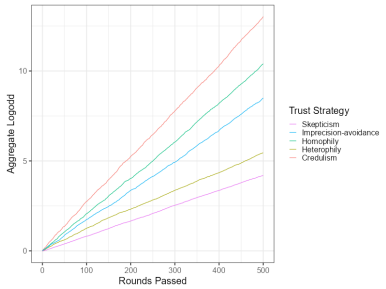
### 4.5.1 Sparse Connectivity

It will be helpful to use an example community to demonstrate the performance of each of the strategies. Consider three agents that sit next to each other on a cyclic network. A is connected to B which is connected to C, but A and C are not connected to each other. Differentiating the strategies can be done by looking at how they deal with the effects of the variance of B's studies. In what follows, suppose that agent B is a particularly imprecise agent. It is important to note that B's evidence has variance even if B was perfectly precise. However increasing B's imprecision increases the variance of their evidence. And this makes the effects more clear. As mentioned above, imprecision makes it more likely that B produces more fluke studies with frequencies further from the true one. Finally, suppose that B starts out with a similar opinion to A, but disagrees significantly with C.

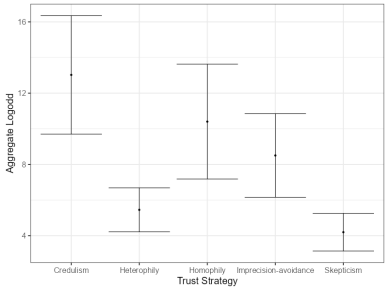
I will start by explaining the effects of the strategies on the diversity of the community. This sets up discussion for both the speed and predictability of the community. Figure 4.1 will facilitate this explanation.

#### *4.5.1.1 Diversity*

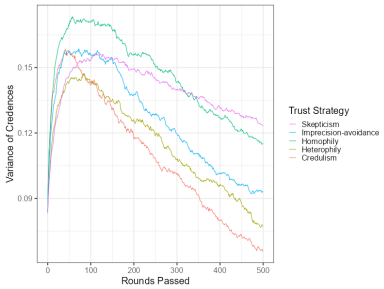
4.5.1.1.1 Homophily Consider first what happens with the homophily strategy, in which agents trust each other the more that they agree. Since B and C disagree significantly, C will discount B's evidence. The greater the disagreement, the more that C will discount B's evidence. Likewise there will be less of an overlap between B and C's pools of evidence. Overlap between the evidence pools entails a correlation in their opinions. The less of an overlap there is between B and C, the less of a correlation there is between their credences. In other words, B and C's evidence are produced by IIDs, identical independent distributions, and




(a) The average aggregate logodd over time




(b) The average aggregate logodd and the standard deviation in the $50^{th}$ round




(c) The variance of credences over time

Figure 4.1 Performance for each trust strategy on a cyclic network with $I = 3$.

when there is no trust, their credences (not merely studies generated) are determined solely by these identical, but independent distributions (ignoring other agents on the network for a moment). When there is trust between the two, their credences are no longer independent (though the research is still generated independently). So the amount of trust dictates to what extent agents have correlation in their opinions. The more that C distrusts B, the less it will be affected by B's imprecision. So when B produces a series of fluke studies, its opinion is moved further from the average, but C is left behind. So homophily leads to agents with disagreeing opinions to have a greater potential for their opinions to diverge (albeit temporarily). This is what I refer to as increasing diversity by *siloing*, i.e. agents having a lack of common information between them. This increases the overall amount of diversity that occurs in the community, and makes it take longer to subside.

In addition, consider the effects that B's imprecision has on A. The closer A's credence is to B's, the less A will discount B's evidence. That means when B produces fluke studies, A and B will both move further from the aggregate opinion of the community. While this decreases the variance between A and B, it increases the variance, not only of B and C discussed above, but A and C. So although A and B's disparity goes down, which decreases the overall amount of diversity, A and C's disparity goes up. In addition, the disparity between B and the rest of the community goes up, as does the disparity between A and the

rest of the community. So while A and B might move close together based on a series of fluke studies, the rest of their cyclic network is blind to it. And the disparity they put between them and the rest of the community out weighs the impact of their own agreement. So homophily not only increases diversity by pulling disagreeing agents apart (siloing), it also increases diversity by pulling agent's neighbors (like A) away from everyone. This is what I call increasing diversity by *extremitization*, agents are pushed further from the aggregate due to evidence that they otherwise would not have heard. So homophily propagates diversity in two ways, siloing and extremetization. This leads it to creating the largest initial jump in the diversity of the community compared to the other strategies (Figure 4.1c).

While siloing and extremetization seem like two sides of the same coin, they are decoupled by both skepticism and credulism as either strategy allows one but not the other.

4.5.1.1.2 Skepticism Skepticism is similar to homophily in that C does not trust B's evidence. Because of that, B's imprecision still does not affect C, and when B produces fluke studies, C is not affected. This means that skepticism allows for the same (temporary) divergence between B and C that homophily does, because their pools of evidence are independent of one another. Hence, skepticism leads to siloing.

The difference is that with skepticism each agent's opinion is independent from *everyone* else's. B's research does not impact A's credence. That means A is not pushed further from C and the rest of the community. So if the goal is to limit the diversity, skepticism has a leg up on homophily by at least eliminating the possibility that agents might extremitize their friends. But because of siloing, it still relies on independent distributions to quell any diversity that arises. This is why on Figure 4.1c, skepticism creates less initial diversity than homophily does. It allows for siloing but not extremetization.

4.5.1.1.3 Credulism A more direct way of 'attacking' the diversity in the community is by making it so that agents fully trust everyone all the time, credulism. This makes it so that any pair of connected agents have correlated opinions (since they are formed by overlapping pools of evidence). So not only does A fully trust B, but C does too. This eliminates the

possibility of siloing. Since B and C's opinions are correlated (and to a lesser extent, so are A and C's), this accelerates the rate at which their opinions converge. Credulism decreases diversity faster than either skepticism or homophily (or any of the other strategies yet to be discussed), because it entails the greatest amount of correlation between agents' credences.

However, that does not mean credulism is the best way to avoid diversity for the entire duration of the simulation. Notice early in the runs, credulism allows a greater amount of diversity than skepticism does (Figure 4.1c). That is because it still allows for cases in which A is influenced by particularly misleading studies of B. Although C is also influenced, the rest of their cyclic community is not. And so even though the community is credulist, B and A move away from the rest of the community increasing the overall amount of diversity. So while credulism does not allow for diversity caused by siloing (independence of evidence), it is still caused by extremitization: A is lead further away from the aggregate than it would without B's evidence. While skepticism allows for siloing without extemitization, credulism allows for extremitization without siloing.

4.5.1.1.4 Heterophily This motivates the idea that to avoid both siloing and extremitization, agents ought to have overlap with those they are distant with (lessening the possibility they move further apart) but also eliminate overlap with those they agree with (which curtails the chance of extremitization). This is precisely what heterophily does. With heterophily, C trusts B's results, and so their disparity is 'actively' worked on by correlating their evidence. So there is no increase in diversity as a result of siloing between B and C (like there is for homophily and skepticism).

On the other hand, A does not trust B's evidence, and so B is unable to extremitize A. Suppose A and B were to initially start together at somewhat of an extreme opinion. There being no overlap means there is no possibility either agent is able to take the other with them any further. It would require the agents to individually produce unlikely evidence for them to continue to move away from the aggregate. Therefore, heterophily curtails the possibility of extremitization as well (unlike credulism and homophily). Since heterophily

avoids both siloing and extemitization, there is less of an initial increase in diversity than any of the alternatives (Figure 4.1c).

The above can be summarized by the following table:

| | extremetization | siloing |
|---|---|---|
| credulism | yes | no |
| skepticism | no | yes |
| homophily | yes | yes |
| heterophily | no | no |

It is worth noting that in the long-run, variance of credences always approaches 0, and so heterophily eventually behaves like skepticism and homophily like credulism. What matters here is what happens in the short term, because that is where we can find interesting differentiation in how these trust strategies navigate the variance of scientific data. In future work however, I plan to consider other (more nuanced) versions of homophily and heterophily that do not collapse to either extreme in the long run.

4.5.1.1.5 Imprecision-Avoidance The above strategies help to demonstrate the role that both siloing and extremitization play in the propagation of diversity by showing different combinations of curbing either effect. The imprecision-avoidance strategy, unlike homophily and heterophily, is not sensitive to the disparity of opinions. Underlying both of siloing and extremetization is the role that the variance of B's studies plays, and that it might push B further from C and take A along with it. By avoiding the extent to which agents are imprecise, the strategy curtails the potential propagation of diversity by lessening (not eliminating) the chance of extremetization. But it also potentially forces siloing. If B is particularly imprecise, this strategy leaves them on their own to achieve more radical credences instead of taming them with more normal studies that they would otherwise have heard from their neighbors. So in this way, the strategy is like skepticism in that it still allows for siloing. In fact, the

more impresice that B is, the stronger the effects of siloing (the less correlation between B and anyone else). So similar to skepticism, it avoids (to some extent) the growth of diversity as a result of extremitization, but does so at the expense of siloing. This is why imprecision-avoidance and skepticism have similar results (Figure 4.1c). The fact that there is some trust with imprecision-avoidance allows for there to be a slightly higher initial increase (because the probability of extremitization is non-zero) than skepticism, as well as a slightly faster decline.

#### *4.5.1.2 Predictability & Speed*

The above tells us how the strategies impact the diversity of the community, but it is important to gauge other metrics too. In chapter 2, I show that promoting diversity by decreasing the correlation between agents' pools of evidence allows for the community as a whole to be more predictable. By making it so the distribution of agents' credences matches the distribution of the evidence, the community opinion is able to 'see' the variance of the evidence. On the other hand, when everyone in the community has correlated opinions, the variance of the data is lost at the community level.

One example of this can be seen by comparing the predictability of the skeptic versus credulist communities.[5] Since skepticism entails there is no correlation between agents' pools of evidence, the fluke studies of one agent do not affect the opinions of the rest. That means for the aggregate opinion to ever go in a particularly extreme direction, it would require a greater number of agents independently generating extreme evidence. Contrarily, when agents have full trust in each other, fewer misleading studies are needed in order for the community as a whole to be swayed in a particular direction. This makes the credulist community less predictable as shown in Figure 4.1b.

Notice also that the credulist community is much faster than the skeptic, and this is because the agents in the former consume more evidence than the latter (Figure 4.1a). This

[5]In Chapter 2 I consider a similar version with a moderate form of skepticism. This more extreme version of skepticism is more helpful here because it is a direct contrast with credulism which facilitates the above discussion.

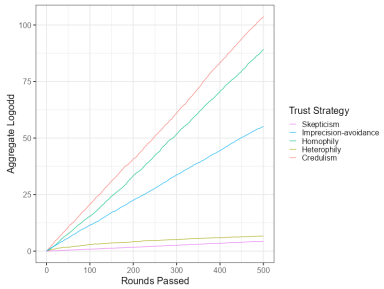
means that when the amount of trust is set at a fixed rate for the entire community, it imposes the same trade-off as the amount of connectivity: More trust means higher, but less predictable opinions and less diversity between those opinions.

The same is not true however for trust strategies which allow for variable amounts of trust, such as homophily and heterophily. With those strategies, the more diverse one, homophily, turns out to be the less predictable one (Figure 4.1b). In previous chapters I have shown how diversifying a community can be a way to allow that community to become more predictable. This shows that heterophily is a strategy for becoming more predictable without having to increase diversity. Heterophily leads to a more unified *and* predictable community. This is because homophily in general, because of extremitization, is more significantly impacted by fluke series of studies. Heterophily on the other hand limits the extent that the community can go in any particular direction as a whole. But it is worth noting that since homophily functionally turns into credulism and heterophily into skepticism, the former turns out to be faster in increasing the aggregate credence (Figure 4.1a).

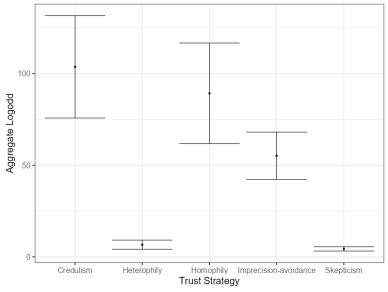
### 4.5.2 Considering More Connections

The above concerns a cyclic network, where every agent has exactly two connections. But that leaves open networks with higher connectivity. Figure 4.2 shows what happens when every agent is connected.

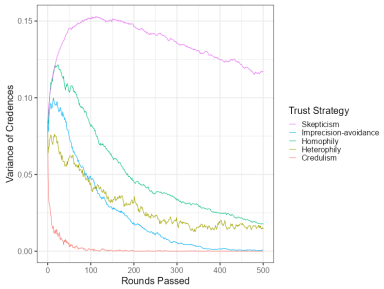
As shown in chapter 1, adding connections, at least on a credulist network, decreases the overall amount of diversity (Figure 4.2c). That is because doing so makes it so that agents have a greater overlap between members of the community. When agents are credulist, more connections guarantees more overlap. As connections increase, extremitization becomes less and less possible for credulists. Suppose that connections are added (evenly) to the network with A, B and C. The more connections that B has, the more of the community will be impacted by B's studies. Likewise, this entails the aggregate itself will be affected by B's studies. So instead of B taking A away from the aggregate credence, B's studies pull the entire aggregate credence closer to their view. This of course happens simultaneously for




(a) The average aggregate logodd over time




(b) The average aggregate logodd and the standard deviation in the $50^{th}$ round




(c) The variance of credences over time

Figure 4.2 Performance for each trust strategy on a complete network with $I = 3$.

all agents (given even distribution of connections) which results in everyone's opinions being closer together. The point is, the more communications there are, the less agents are able to extremitize each other. When there is complete connectivity it is impossible. Further, this decreases the potential for siloing between *any* two agents. Credulism already eliminated the possibility of siloing between connected agents, so when connections are made maximal, siloing becomes impossible. Therefore, with credulism on a complete network, both causes of diversity are silenced. Any diversity that the agents start with is quickly washed away.

In contrast, when there is the potential for agents to distrust each other, increasing connections does not necessarily decrease diversity. Skepticism is an easy example of this, because the strategy makes the network (and amount of connections) completely irrelevant. Therefore, siloing is always a danger for skepticism, and extremetization never is.

But a more interesting example is that of homophily, where connections between agents do make a difference. Notice that even when the network is complete, there is still a significant amount of diversity in the community (Figure 4.2c). That is because agents that happen to have disparate views discount each other's evidence. That makes it more likely that either camp produces more studies that point them in a direction which the other does not see. This allows the diversity of the community the ability to persist much longer than it would if there was complete overlap. This means that homophily, like skepticism, allows for an increase

in diversity even on a complete network. Adding connections does help in the sense that it increases the amount of total evidence updated on which in turn will accelerate the speed at which they approach a credence of 1 (and lose diversity). But the greater connections do not provide any overlap before agents agree, and so does not directly 'attack' the diversity in the way described above — it still allows for siloing and even extremitization.

On the other hand, heterophily is affected by the amount of connections in that it allows more opportunity to directly undermine the production of diversity (unlike homophily and skepticism). That is, each agent is able to correlate their evidence with *everyone* on the network to the extent they disagree. To the extent they agree, they discount each other's evidence. This makes it so siloing and extremitization are minimized. This means the community does not increase its diversity, but decreases much like the credulist network. The difference between the two is that as agents get close to each other, they trust each other less. This means the rate at which they converge, while always positive, is always slowing.

But notice what all of this does to the speed at which the community aggregate increases (Figure 4.2a), and how predictably it does so (Figure 4.2b). Heterophily turns out to be much more like skepticism, in that the community aggregate is low, yet predictable. So while the two differ significantly in the amount of diversity they induce, they produce the similar aggregate performances. Meanwhile credulism and homophily both achieve high, less predictable confidence levels, because both communities are more susceptible to the fluke studies of their agents. This is largely because the potential for extremitization that is present for both (in low connectivity) is manifested as unpredictability when there are many connections. The fluke studies do not just move B and A, but large parts of the (or the entire) community.

These results can be summarized in the following table (notice how this lines up with the previous table):

As for imprecision-avoidance, it has some initial increase in diversity because it still allows

| | higher, less predictable credences | initial increase in variance of credences |
|---|---|---|
| credulism | yes | no |
| skepticism | no | yes |
| homophily | yes | yes |
| heterophily | no | no |

for siloing. However, because it has less of a potential for extremitization, it has less of an initial increase than homophily. Since they both use some correlation between agents, they approach 0 diversity quicker than skepticism.

### 4.6 Conclusion

Three conclusions can be made based on the above.

First, the diversity of opinion can be caused by the variance of scientific evidence in two importantly distinct ways. Often times diversity of opinion, as well as both siloing and extremitization are thought to be the result of malfeasance or ineptitude, but this shows how diversity arises purely out of the nature of science itself. (Remember that while imprecision increases the variance of evidence making the above effects more apparent and imprecision-avoidance worth considering, there is always variance to sampling.) Further, the way that variance of evidence causes this is surprising. Because it is not only in one way, but two. This means that other dynamics might complicate one of these ways independent of the other. Understanding these features of the nature of scientific research is a vital first step in understanding the social enterprise of science.

Second, studying epistemic communities requires *variance measures*. None of the analysis above concerning the diversity of the community or its predictability would be possible by only averaging the data of the community. Variance measures, like average measures, are an indispensable part of understanding the dynamics of a community. This opens the door up for further research concerning essentially every formal, social epistemology project. All but a slim minority consider anything except for aggregate opinion, and so there is likely insight

to gain in just about every one by considering the variance within the community of a single run (diversity) as well as the variance between individual runs (predictability).

Finally, heterophily deserves significant future investigation on par with what homophily has received. Homophily seems to have gotten its foot in the door of people's minds because it is plausibly what people actually do in many cases. However, as epistemologists, we ought to be concerned with what we *should* do. This chapter shows what heterophily can do, and as a result opens up our repository of options. This option in particular curtails both siloing and extemetization without causing unpredictability.

# 5. MITIGATING & EXACERBATING BIAS-INDUCED ECHO-CHAMBERS WITH TRUST

## 5.1 Overview

In this chapter, I show that heterophily, a strategy in which agents trust each other less when they agree more, can help mitigate severe challenges imposed by biased agents. I provide a novel characterization of echo-chambers, entrenchment, and group-think. This involves echo-chambers which consist of feedback-loops, as opposed to the siloing away of information. These echo-chambers can manifest as either entrenchment (a community forever split on an issue) or group-think (a community unanimous on an arbitrary answer) depending on other conditions such as the amount of connectivity. I compare heterophily with four other strategies: credulism (full trust), skepticism (no trust), bias-avoidance (trusting others less to the extent they are biased) and homophily. This final strategy is one where agents trust each other more when they agree more, and versions of it have been explored demonstrating both good and bad effects. This chapter has three upshots: i) It explains how a very simple, yet real form of bias can affect communities. ii) It provides a new understanding of echo-chambers (which does not depend on silos) as well as the relationships between them, entrenchment, and group-think. iii) It shows how heterophily can mitigate the effects of echo-chambers, and likewise establishes the strategy as worthwhile for further research.

## 5.2 Introduction

This project poses a problem, the effects of bias on researchers, and a potential solution to that problem, trusting others to the extent one disagrees with them.

Bias among scientists poses significant a threat to the chances that they come to have accurate opinions. For instance, the data that a researcher produces might be (intentionally or unintentionally) altered in a way that supports their opinion, thereby undermining the ability of that data to aid in forming accurate opinions. In this chapter, I show how this sort of bias can lead to feedback-loops between members. These feedback-loops can be used to characterize a version of echo-chambers which does not depend on being excluded from certain sources of information. These echo-chambers can manifest in two different ways: The community can become forever split on the issue (*entrenched*), or they can become unanimous on an arbitrary answer (*group-think*). Hence this project shows how these two seemingly unrelated challenges for a community can have a fundamentally common cause: feedback-loops. In either case, the chance that any given agent's opinion is accurate is largely a matter of luck.

Trust strategies are plausible ways in which agents might try to mitigate the effects of bias on their opinions. By discounting the evidence that one hears based on who shared it, agents might hope avoid the impacts of bias. I consider multiple strategies for agents doing so. Most trust strategies I consider (which have precedent elsewhere in the literature) either allow the community to fall into one of the two cases above, entrenchment or group-think. That is, by avoiding entrenchment, strategies force arbitrary acceptance of some answer. By avoiding blind unanimity, agents slip into an irresolvable division in their community. However one strategy, *heterophily*, which has not been so far discussed in the literature (aside from the previous chapter), shows promise. With this strategy, agents are able to (at least partially) undermine the formation of echo-chambers and thereby mitigate the effects of bias. As a result, heterophilic communities are at least sometimes able to achieve non-lucky opinions that become more accurate over time.

I make these findings using a computer model developed in Chapters 1, 2 and 3. The model concerns a network of researchers who statistically generate studies to form their opinions. The point is to focus on a very simplified but core aspect of much of scientific research. By keeping the mechanisms in the model to a minimum, what is illustrated here applies to any domain that concerns sharing statistically generated research distorted by the version of bias under consideration.

This chapter has three upshots: i) It explains how a very simple, yet real form of bias can affect communities. ii) It provides a new understanding of echo-chambers (which concerns feedback-loops and not silos) as well as the relationships between them, entrenchment, and group-think. iii) It shows how heterophily can mitigate the effects of echo-chambers, and likewise establishes the strategy as worthwhile for further research.

In the next section I give some of the background literature concerning bias, echo-chambers, and trust strategies. This is followed by details of the model itself. The results section is broken into three parts: First, I explain the impacts of bias when agents fully trust one another. Second, I explain the ways in which three other trust strategies interact with bias. Third, I show how heterophily at least partially mitigates the effects.

## 5.3 Background

### 5.3.1 Bias

In "Why Do Scientists Lie?" (2021), Bright begins with the case of Brian Wansink. Wansink was a nutritional scientists who fabricated his data in order to support his claims. In the words of Bright,

> He had authored 'studies suggesting people who grocery shop hungry buy more calories; that preordering lunch can help you choose healthier food; and that serving people out of large bowls encourage them to serve themselves larger portions', and on the basis of his expertise on these phenomena had been feted by the press and called to assist Google and the US Army run programmes designed

> to encourage healthy eating. The problem, however, was that he had not in fact produced evidence for these claims. On some occasions he had misreported his own data – that is to say, lied about what he had found. And on other occasions had more or less tortured his data with no regard to proper statistical method so as to give him something, anything, to publish.

This is the sort of distortion of evidence that I am concerned about — specifically where Wansink started with some presumably good data and distorted it in order to fit his opinion. Bias like this can occur in the world in a variety of different ways depending on the field. For instance, a medical researcher might inflate the number of successful patients in their trial. Some social scientists might be incentivized to distort their data because null studies are far less published. Or, various industries can provide monetary incentives for researchers whose studies happen to say the right thing. The question is not whether or not any of these cases are justifiable, the question is simply what should a community do in their presence. Philosophy of science and science itself would both be easier (not easy though[1]) if bad actors did not exist. But that is not the case, and it is important that epistemologists address this.

But one does not need to think of only clear cases of deliberate fraud for this conception of bias to apply. For instance, consider a study where there is some borderline case: it is not immediately clear if it is a success or a failure. Perhaps in a medical case, it is not clear if a patient truly recovered from their ailment. In such a case, a biased agent might *subconsciously* interpret their observation in a way that supports their view more often than they should (i.e. than what matches reality). In this case, the effect of the agent's subconscious bias is the same as the conscious bias from above: some sample has been taken, but its frequency is not a perfect reflection of what is intended to be captured. It is instead distorted to some extent due to the agent's level of bias. The upshot of this is that bias of this type —distorting information to support one's opinion— is a simple, yet crucial concern to consider.

[1]The previous three chapters demonstrate this is the case. Scientific choices are far from trivial, even when only the inherent variance of data sampling is involved.

### 5.3.2 Echo-Chambers

Echo-chambers have begun to receive a significant amount of attention due to their high prevalence in real-world communities. However there are many elements which might constitute real world echo-chambers from 'purely epistemic' dynamics to ones about social or economic divisions. The job of philosophy is to parse these out. In particular, the version of bias just discussed plays a crucial role in echo-chambers independent of other factors. One might think that non-ideal concerns such as bias and echo-chambers are too complicated to learn about through simplified models such as this. That sort of thinking is exactly what this project undermines. I show that even a very simple version of bias can have drastic implications with regard to echo-chambers. The ways that these effects interact with other features of the model is highly informative. The upshot is that we begin to untangle the chaotic knots of epistemic problems communities face. By isolating the dynamics of this particular form of echo-chambers, and understanding its various complications, we are in better situation to understand other, more complicated dynamics (whether in more complex models or the real world).

Nguyen, 2020 contributes to this process by articulating echo-chambers as epistemic bubbles. That is, agents do not have access to all of the information in the community. This is what I refer to in Chapter 3 as *siloing*. In these cases, the communities face certain dynamics because not all agents are exposed to all information. This is a crucial part of real world echo-chambers in which agents can be, whether intentionally or not, isolated from certain streams of evidence.

But siloing (or epistemic bubbles) only tells part of the story. It is important to distinguish the effects of siloing from other features of echo-chambers. In particular, I am concerned with the *feedback-loops* between its members. Two agents can mutually intensify each other's opinions if the increase of one of their opinions always leads to an increase in the other. This conception does not depend on there being a lack of contrary information, but instead an overabundance of confirming information. So these feedback-loops (and likewise

echo-chambers) can occur even when all agents have access to all the information produced.

This means that echo-chambers, constituted purely by feedback-loops and setting aside siloing, deserve close attention. Before working out the complicated dynamics of real-world echo-chambers, it is necessary to first understand their fundamental aspects.

### 5.3.3 Trust

In Chapter 3, I consider how various trust strategies perform when members of a community are to some degree imprecise in their data generation (though they are not biased). Doing so showed how the various trust strategies navigate the various levels of imprecision in regards to how much diversity in the community they allow for and how predictable the aggregate opinion of the community is. That project is similar to K. J. S. Zollman, 2015, where Zollman also uses a form of imprecision (what he calls 'inherent unreliability') to motivate the use of trust strategies. Chapter 3 builds on Zollman's project by considering more metrics: *diversity* and *predictability*. The diversity of the community is the spread between (variance of) the individual opinions in the community. The predictability is the variance in the aggregate performance of the community from one run to the next. These turn out to be indispensable for understanding the dynamics of the community and are even more important here. Chapter 3 also considers an additional trust strategy, *heterophily*. Heterophily is a strategy where agents trust others the more that they disagree with them. This contrasts with *homophily*, where agents trust others the more they agree. Homophily has received considerable attention for both its strengths and weaknesses, while heterophily has not been studied before Chapter 3. In that chapter, I show that for the other strategies, being more predictable means being more diverse. Heterophily on the other hand allows for the least amount of diversity in the community while being the most predictable in aggregate.

This project continues the same idea as that of Zollman and Chapter 3: trust strategies are tested with regard to how well they deal with each individual's unreliability. However in this chapter, the problem is more severe for a community. Since the agents are not merely imprecise, but biased. Since heterophily has shown promise before in keeping down diversity

while not yielding an arbitrary answer as the result, it is reasonable to think it might perform interestingly here too.

I test five trust strategies in total:

***Credulism***:

Agents trust each other fully.

***Skepticism***:

Agents do not trust each other at all.

***Bias-avoidance***:

Agents trust each other less the more biased they are.

***Homophily***:

Agents trust each other in direct proportion with the extent that they agree.

***Heterophily***:

Agents trust each other in direct proportion with the extent that they disagree.

## 5.4 Model

The model concerns a network of agents that share statistically generated evidence to update their opinions concerning a particular problem. The value of this model is that it applies to the broad range of scientific communities that rely on random sampling, and it depends on minimal epistemic assumptions. I will first explain the basics of the model before generalizing it to account for bias and then trust. Subsequently, I explain the metrics used to explore the model.

The problem in question can be thought of as whether or not a particular frequency in nature is above or below a given threshold. One can think of this as doctors determining if a medical treatment's success rate is higher than a known standard, but (as I discuss in Chapter 1), there are many interpretations of the model. All that is important is that the agents use statistically generated evidence to update their opinions. The community consists of $N$ many agents that each produce $n$ new data points every round. The observed frequency

of a study is $\frac{k}{n}$ where $k$ denotes the number of successes for the study. The $\frac{k}{n}$ of any given study is likely above or below the true frequency, but overtime the aggregate frequency of the studies reflects it. Agents research every round. This entails the amount of evidence that is produced each round is constant, which ensures that the dynamics that unfold are due purely to the flow of information and not how much is produced.

Each agent has a credence indicating the confidence in their opinion that the frequency is above the threshold (that the new treatment of the disease is better). Increasing their accuracy means increasing their credence, since the new treatment is in fact better. I sometimes represent the opinions of agents by applying a logarithmic transformation to their credences, the result of which I refer to as the *logodd* of the agent (the result represents a logarithm of the odds ratio). This doesn't require any extra philosophical assumptions and merely provides another useful lens through which to analyze the model. For details, refer to Chapter 1.

Agents share their studies with those they are connected to on a network. In chapter 1, I show that increasing the amount of connectivity can reduce the diversity of opinions, and this turns out to have significant effects on the impact of bias below. I demonstrate my results with two networks. In future work, I plan to investigate the effects of bias and trust on more network structures, but looking at the complexities which result for network connectivity alone are sufficient for a deeper understanding of echo-chambers, entrenchment, and group-think. The effects of network connectivity can be demonstrated with the *cyclic* and *complete* networks. The cyclic network is one where every agent only has two neighbors, and no two neighbors share a third (assuming the population is $> 3$). The complete network has all agents connected to each other.

Bias can be modelled by giving each agent their own *bias value* which indicates by how much an agent might distort their evidence in favor of their view. When an agent, $j$, produces a new sample of data, a number randomly generated in accordance with their level of bias, $b_j$, is added to the number of successes they observe, $k$. $b_j \leq B \leq 1$, where $B$ is the maximum

level of bias for members of the community. More specifically, $k$ is reassigned in the following way:

$$k_{biased} \leftarrow k_{original} \oplus |randnorm(0, b_j)| * 2(c_j - .5) * n$$

The function $randnorm(m, v)$ yields the nearest integer to a real number generated with mean $m$ and variance $v$. $x \oplus y$ is the sum of $x$ and $y$, but rounded to 0 if it is less than 0 or to $n$ if it is greater than $n$ (the number of trials in the study). Hence $randnorm(0, b_j)$ generates a number based on a distribution centered around 0 with a variance $b_j$ — which indicates the level of bias for that agent. The absolute of that value (which is almost always less than 1, because $B$ is less than 1) is multiplied by the size of the samples taken in order to make the distortion of an agents' bias relative to the size of the tests taken.[2] Notice that an agent's bias can only cause their results to be distorted in a way that conforms to their credence. That is, if $c_j \leq .5$, then the bias will skew the study down. Otherwise it will be skewed up. The closer that $c_j$ is to 0 or 1, the more impact the bias has. In addition, the greater $b_j$ is (the more inherently biased the agent is), the more likely that any given distortion will be larger. Note that the results from Chapters 1, 2, and 3 (as well as Chapter 5 to come) concern perfectly unbiased communities, i.e. $B = 0$.

Trust can be modelled by having agents discount the evidence they hear from their connections. Each agent $j$ assigns a weight between 0 and 1 for each connection $l$ that they have, denoted $w_{jl}$. Before updating on the evidence they hear, an agent multiplies both $k$ and $n$ by the $w$ they have for that agent. Hence, the level of trust does not change the frequency that the study indicates ($\frac{k}{n} = \frac{w*k}{w*n}$), but it does decrease the level of impact of that study on the agent's opinion. (Note a $k = 60, n = 100$ study impacts an agent's opinion more than a $k = 6, n = 10$ one.) This allows us to redefine the above strategies in terms of

[2]One might wonder whether this is appropriate, since larger studies might be thought to wash out the effects of an agent's bias. Hence one might think that the effects of bias should not scale with the size of the tests. But this is wrong. If an agent inflates their numbers or subconsciously misjudges cases — all for individual data points —, then the potential for these same issues would be present for every particular patient or data point considered. The bias is a danger for every data point, and likewise the effects of bias scale with the size of the tests.

$w_{jl}$.

***Credulism***:

$w_{jl} = 1$

***Skepticism***:

$w_{jl} = 0$

***Bias-avoidance***:

$w_{jl} = 1 - |b_l - I|$

***Homophily***:

$w_{jl} = 1 - |c_j - c_l|$

***Heterophily***:

$w_{jl} = |c_j - c_l|$

## 5.5 Results

Figures demonstrating the results can be found in the Additional Materials section. The results are broken into three sections. First, I explain in the context of credulism how bias can lead to feedback loops, based on which a novel characterization of echo-chambers, entrenchment, and group-think is provided. Second, I explain how the other trust strategies, save heterophily, interact with these dynamics. Third, I explain how heterophily partially mitigates the effects. This section is followed by a conclusion of the findings.

### 5.5.1 Echo-Chambers, Entrenchment, & Group-Think (Under Credulism)

Consider a scenario where two agents, A and B, are connected to each other on a network and have credences above 0.5. In this situation, the evidence they produce will be distorted in a way that makes the treatment seem more successful, and as a result, their credences will increase more than they otherwise would (without distortion). Each agent intensifies both their own credence and that of their peer. This is visualized in Figure 5.1. The diagonal lines from either agent indicates the effect of their sharing on their peer. A's sharing makes it so that B is more likely to be more confident in the same answer in the next round. And

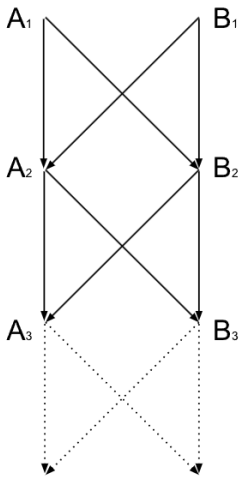



Figure 5.1 Feedback loop between agents A and B over 3 rounds (indicated by the subscript). The arrows indicate the impact of each other's evidence on each other and themselves: A's bias makes it more likely that B will share biased information in the following round (increasing A's credence), in addition to the effect that A's bias has on A in the next round. Meanwhile B's bias makes it more likely that A will do the same.

this makes it more likely that B will share even more biased evidence (since the distortion increases with their level of confidence). So A's sharing to B has an indirect effect on A in the subsequent round. The vertical lines represent the agent's impacts on their own credences. Each round, the agent's biased sample makes it more likely that their credence will be more intense in the following round. This in turn makes their bias even more impactful. Both the effect that the agents have on themselves and on each other are instances of *feedback-loops*. The increasing of an opinion leads to another increasing of opinion, *ad infinitum*. Agents are in a feedback-loop when their opinions mutually intensify each other. This is the central feature of how I characterize *echo-chambers*.

It is important to note that echo-chambers can exist even in the presence of contrary evidence. Agents can be connected to other agents who hold opposing views, but if they hear more evidence from those they agree with than from those they do not, the feedback loop will persist. Suppose B is also connected to C, where C has a credence below .5. The

effect of A's and B's biased studies (controlling for their respective $b$ values) overpowers that of C's studies.

The formation of an echo-chamber that drives toward either extreme (0 or 1) depends on the agents' level of bias and the difficulty of the problem at hand. If the bias is strong enough to make any study indicate a frequency above a certain threshold, and the observed frequencies are so high above that threshold, an echo-chamber can form. This means that bias needs to meet some threshold in order for it to cause echo-chambers. In future work I plan to investigate exactly what this value is and how it relates to other parameters. But it is important to note that nothing hinges on the particular values used, since these models are not meant to be tuned to real world scenarios (along with any models in this literature). What is important are the dynamics themselves. For this reason, for the results below I focus only on cases where there is a sufficient level of bias for echo-chambers. This allows us to see in what ways they manifest and how trust strategies interact with them.

Since echo-chambers can persist in the presence of contrary evidence, this makes it possible that two echo-chambers of disagreeing opinions can exist within the same community. Regardless of how many members of either echo-chamber are connected, each individual member hears more evidence in support of their respective view than against it. For instance, suppose agent D is connected to A and C on the network but not B, and suppose D has a similar opinion to C (and differs from A and B). This means that each agent will (independently) hear two studies in support of their view and only one against. A from B and D; B from A and C; C from B and D; D from A and C. In this situation, the community will be endlessly split. This is what I call *entrenchment*.

This gives an early indication for how the network connectivity impacts the community. Consider adding connections so that A is connected with C and B is connected with D (so that every member is connected to every other member). In this case, every member will receive exactly the same studies. This means there will quickly be no difference in any of the agents' opinions (because they will all be formed on exactly the same evidence). However the

direction that the community ends in is determined by which initial credences were paired with which levels of inherent bias. The stronger combination will convince the community quickly and forever. This is what I call *group-think*: the community has unanimously chosen an arbitrary answer and will spend forever (baselessly) reassuring themselves it is correct.

Notice that in each of these cases, the opinions of the members hardly seem justified — even if they happen to support the correct opinion. In these cases, the opinions are not safe in the sense that they could have just as easily been wrong. They are not sensitive to evidence, because after only a few rounds of the simulation, it is determined the direction members will go forever. This might sound like too dismal a setting to test trust strategies — that if a community is plagued with a form of bias like this, then there is essentially no chance of them finding the truth. One might think the deck has been stacked against the community.

Below I show that this is not the case. Heterophily is able to provide a promising start to alleviating the echo-chambers. But first, it is instructive to see how the other strategies fare.

### 5.5.2 Other Strategies

The above concerns how credulist agents behave in the midst of bias, but one might wonder whether agents discounting each other's evidence in various ways might allow them to navigate each other's bias to give them a plausible chance of a justified opinion. Credulism requires that feedback loops emerge, because it makes it so that one agent increasing their credence causes their connections to be pushed in the same direction. One might think moving to skepticism solves this problem.

Skepticism partially solves the problem, but not entirely. Skeptic agents are unable to form echo-chambers with those of like minds. However, skeptics are still affected by their own bias. Likewise, they are condemned to create their own singleton echo-chambers which move in whatever direction is decided by their (random) initial credence, ignoring the rest of the community. Consider Figure 5.1 but imagine the diagonal lines are missing. In

this case, agents' bias still affects themselves but not those around them. This results in an intensely split community, since the agents will move in whatever direction their initial credence indicated without ever hearing from their peers. However the agents have a less intense opinion than they would have given credulism, because they have less evidence to update on. Likewise the aggregate logodd of the community becomes very predictable, since in every simulation the community is split around modest opinions, leaving the average only just above 0. Also note, skepticism makes the network structure irrelevant. So entrenchment happens for skeptic communities no matter the network.

What about bias-avoidance? This strategy might be thought to attack the problem at its source. Agents are distrusted the more that they are biased. But while this strategy might discount biased evidence, it does not reverse the impact of bias in a way that makes the evidence itself unbiased again. Any particular evidence that one gets would still be skewed in a particular direction, though potentially less so. This means that avoiding bias allows agents to be less affected by the most extreme (with regard to their inherent level of bias) members of the community. Nonetheless, their opinions are dictated by only biased evidence in the same way credulist opinions are, albeit with toned down impacts of bias. Therefore, bias-avoiding communities behave similar to, but less extreme than, credulist communities. With sparse connectivity, they entrench, while in heavily connected communities they engage in group-think. In either case, the resulting opinions are less strong than they would be under credulism, due to less of an impact from bias. But the feedback loops still emerge, albeit in a less severe way. This strategy highlights that a fixed rate of discount, even one that is attached to an agent's inherent level of bias, does not undermine the echo-chambers. That is, when agents hold constant the level of trust they have for other agents in the simulation, an equilibrium will quickly emerge just like in the credulist cases. And this leaves agents to go monotonically toward 0 or 1 in every round of the simulation. In order to stop this from happening, we must consider trust strategies that vary the amount of trust as time goes. Doing so in the right way may allow agents to break out of echo-chambers or curtail their

powers from intensifying.

One way that agents' trust levels can be made variable is by tying them to the opinions of agents. Homophily increases the amount of trust between agents when they have more similar credences. This means that as agents, such as A and B, share biased evidence with each other that brings them closer together, the effects of their evidence (setting aside that their credences are getting stronger too) have a stronger impact on either agent. Meanwhile, the evidence from agents that one disagrees with becomes virtually ignored. Hence homophily exacerbates the effects of bias and increases the intensity of the echo-chambers. The amount of trust within echo-chambers continues to increase, while the trust between members of opposing echo-chambers disappears. So regardless of the amount of connectivity, even in complete connectivity, homophily leads to entrenchment. This is similar to skepticism which also leads to entrenchment always (given sufficient bias), but the difference is the extreme opinions of the homophilic community compared to the skeptic community. While skepticism leads to entrenchment of relatively moderate opinions, the greater information open to the homophilics leads to entrenchment with much stronger opinions. While homophily does nothing but amplify the effects of bias, it suggests a promising way forward: its non-evil twin, heterophily.

### 5.5.3 Heterophily

While homophily exacerbates the strength of the echo-chambers by intensifying the feedback-loops and eliminating countering evidence, it is not difficult to see how heterophily does the exact opposite. If A and B are heterophilic, then the more they agree, the less that they will consider each other's evidence. While early on the studies make at least some impact, this diminishes the more the agents agree. This directly curtails the propagation of the feedback loop: as it goes on, it gets weaker. On the other hand, consider the effects of C's evidence on B. The further that C's opinion gets from B's, the more weight that B will give to it. This means C's evidence (given that it is biased in the direction away from B's echo-chamber) actively pulls B away from their echo-chamber the longer it goes (or stops it

from forming in the first place). So in both of these ways — decoupling A's and B's evidence, and coupling B's and C's —, heterophily curtails the effects of feedback-loops.

On a cyclic network, notice that heterophily allows for the least amount of variance compared to any of the alternatives considered. Further, the amount of variance actually decreases, albeit extremely slowly. This is because when there are two entrenched echo-chambers on a cyclic network, there must be at least two agents in either echo-chamber that sit at the 'edges' of the echo-chambers. These agents begin to trust the agents in the opposing echo-chamber more than the ones in their own, because of the disparity/similarity in the credences. While the edge agents take their own biased results at full value, they take each other's evidence at increasingly stronger value too (limiting at full value, same as their own). This means on cyclic networks, there is the chance for the echo-chambers to be dissolved away at the edges. At least one of the edge agents ends up leaving their echo-chamber to become an *intermediate* agent. Intermediacy characterizes not only their position between echo-chambers, but the agents' opinions. Intermediate agents like these are affected by the biased evidence on either side, which can essentially cancel each other out. This results in a credence that sits in between. This means that agent's own bias has less of an impact, since the effects of bias increase with the intensity of opinion. So an intermediate agent allows for the generation of what is potentially the least biased evidence in the community. These intermediate agents are quite important, because they are the only agents not to monotonically approach one answer or the other. Their opinions might even be called properly sensitive to evidence. Further, a cascade of intermediate agents converting other edge agents can potentially dissolve the entrenched echo-chambers. The upshot is that the echo-chambers have the potential to dissolve in a way that was not possible for any of the other strategies. It is not yet clear if this *always* happens, but heterophily is interesting because it *possible* in a way that no other strategy does.

For a complete network, everyone is exposed to the same evidence, but each agent discounts it in a slightly different way. Even still, the community quickly converges to a shared

credence. That is because to the extent that anyone is far from that credence, they accept the evidence of those that are close (evidence which is in fact responsible for those members being close to the average). That means disagreeing agents might stray temporarily but will be pulled back to the community average just as quickly as they stray. When all of the agents have landed on a similar side, they are skeptical of each other and only rely on their own individually produced evidence to push them. In this way, the agents still engage in singleton echo-chambers including only themselves in isolation. As their credence intensifies, so does the impact of the bias in making that credence more intense. But notice, these echo-chambers are less intense than the group-think echo-chamber that credulism yields, or the entrenched echo-chambers of homophily. Instead, the community's opinions have the intensity of a skeptical community (where opinions are based on less information so have less chance to get extreme). The difference between heterophily and skepticism in this case comes down to one and not the other allowing for diversity, since heterophily requires that all agents move to the same side. A summary of these results for the complete network are captured in the following table:

| | group-think or entrenchment | lower or higher aggregate logodd |
|---|---|---|
| credulism | group-think | higher |
| skepticism | entrenchment | lower |
| homophily | entrenchment | higher |
| heterophily | group-think | lower |

## 5.6 Conclusion

There are three main upshots to the above.

First, this provides an important isolated analysis of a certain simple, yet present and powerful form of bias. Bias like this takes many forms in the real world, both as deliberate actions and subconscious mistakes. This model provides a simplified account of such bias,

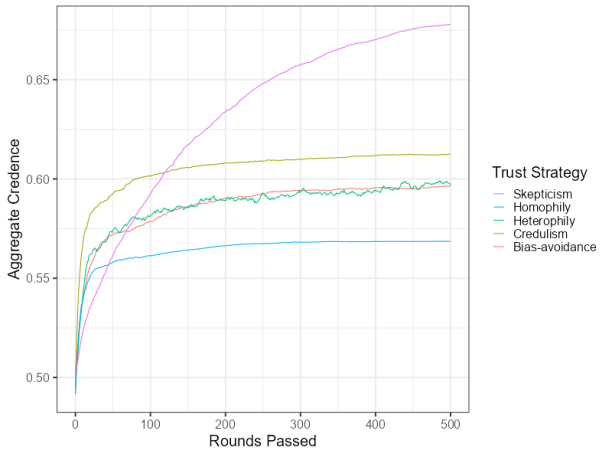
and in doing so sets up future research in multiple ways. i) The level of bias turns out to have very interesting properties which I have already begun investigating. Instead of having continuous effects, there are thresholds beyond which the possibility of echo-chambers (or their resiliency) 'jumps'. Similar to $H_2O$ behaving differently above or below $0°C$, the community experiences a phase change depending on the level of bias. ii) There are other complications to the network, such as other parameters of networks which presumably have very interesting interactions with bias just as connectivity does. Here, we can observe that connectivity changes the way bias manifests (discussed in a moment), and the same is likely true for other features. iii) Having one version of bias makes it easier to conceptualize and model distinct forms of bias, which now can be considered and compared to this one. This form does not capture everything, but it is important to start somewhere.

Second, echo-chambers can be constituted without regard to silos, but feedback-loops alone. The issue is not simply that the members have blocked out information, but that they have over-inflated the impact of other information. With this insight, we can see a similarity between entrenchment and group-think. Both concern subsets of the community where agents' beliefs are arbitrarily determined. The difference is simply the existence of others that happened to believe differently. In this way, we can see how siloing and feedback-loops interact. The feedback-loops causes the echo-chambers to exist, and the siloing determines how many echo-chambers there will be. This provides a deeper understanding of real world echo-chambers.

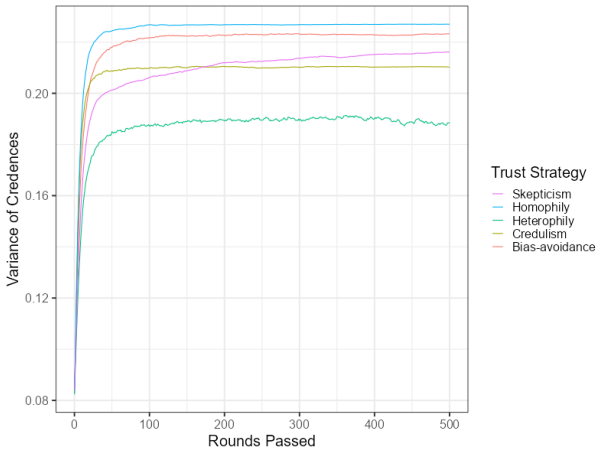
Finally, avoiding arbitrary opinions requires more than simply exposure to contrary evidence, but an active method of seeking balance between the evidence. The results for credulism shows that both entrenchment and group think can occur, even when one is willing to accept the evidence of those they disagree with. The way that heterophily is able to combat feedback loops is by not only seeking dissenting information, nor by avoiding all others entirely, but by both avoiding confirming information and seeking out dissenting information. This sets up future work to investigate in what other places heterophily is able

to mitigate echo-chambers.

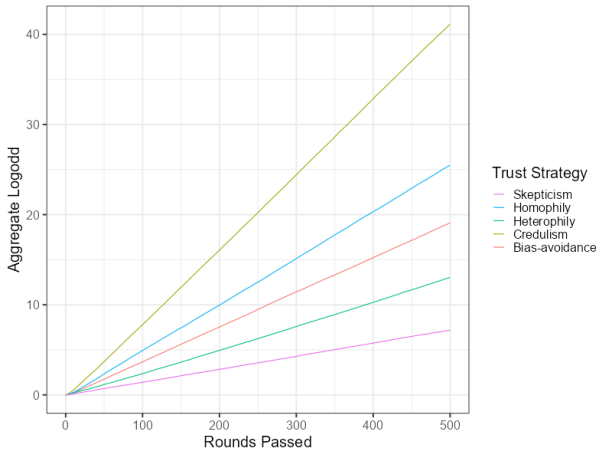
## 5.7 Additional Material



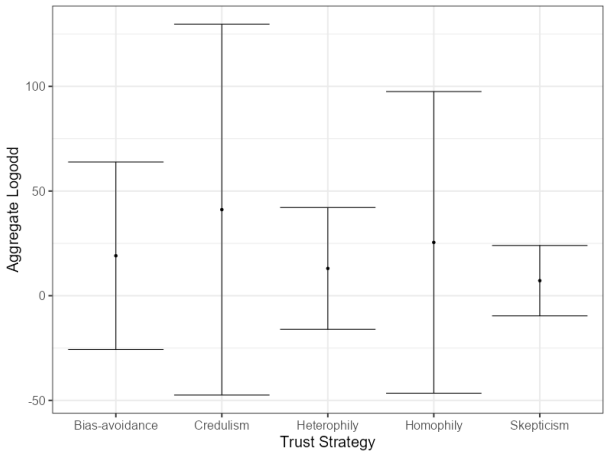

(a) The average aggregate credence over time



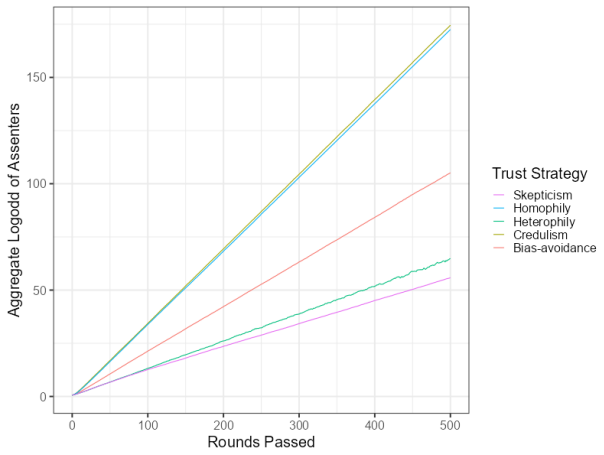

(b) The average variance of credences over time



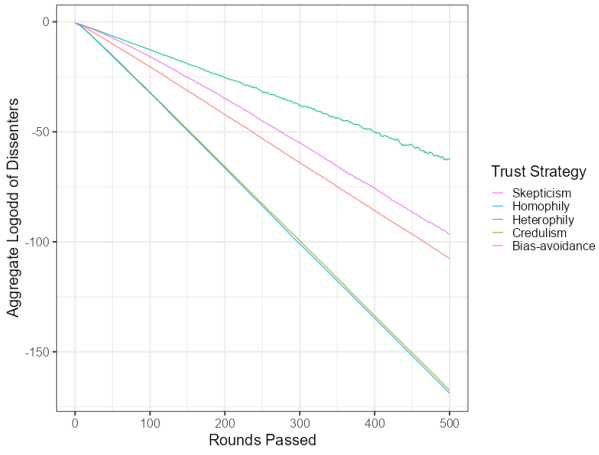

(c) The average aggregate logodd over time




(d) The average and standard deviation of the aggregate logodd after the $1000^{th}$ round of the simulation.

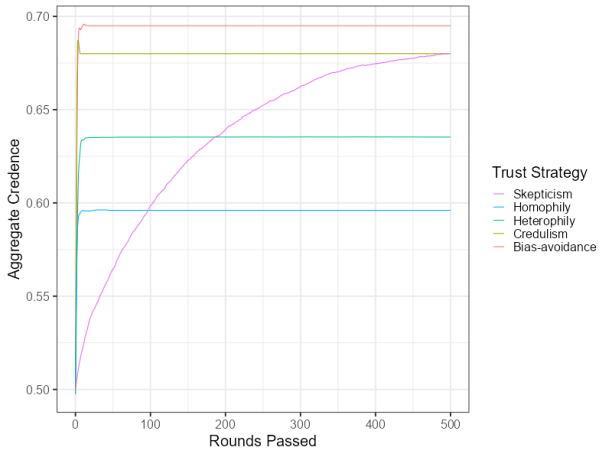



(e) The average aggregate logodd of the dissenters (agents with credence $> .5$) over time

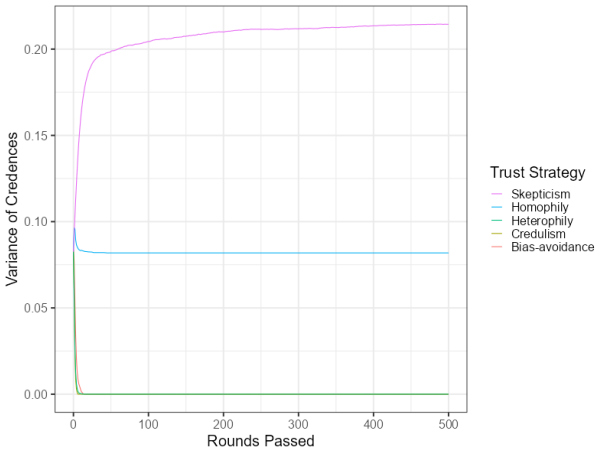



(f) The average aggregate logodd of the dissenters (agents with credence below $\leq .5$) over time

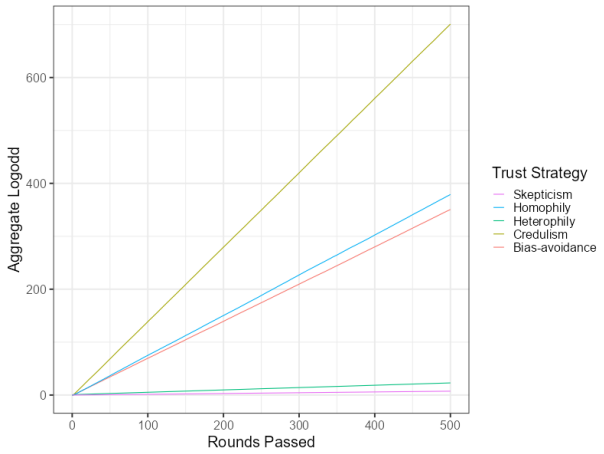
Figure 5.2 The performance of trust strategies on a cyclic network with biased agents



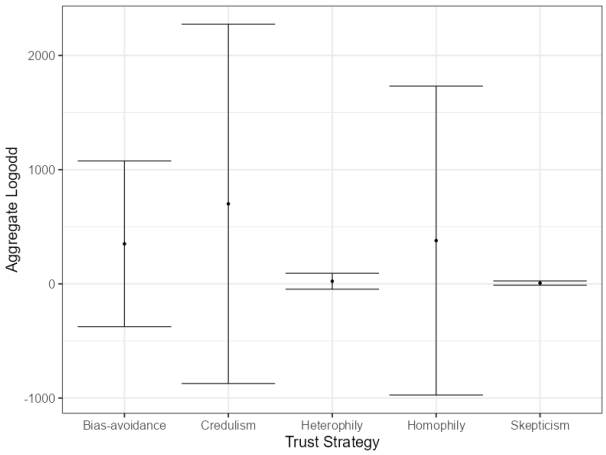

(a) The average aggregate credence over time



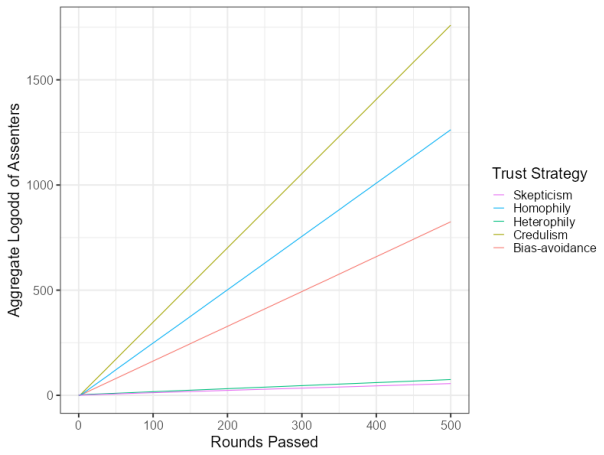

(b) The average variance of credences over time



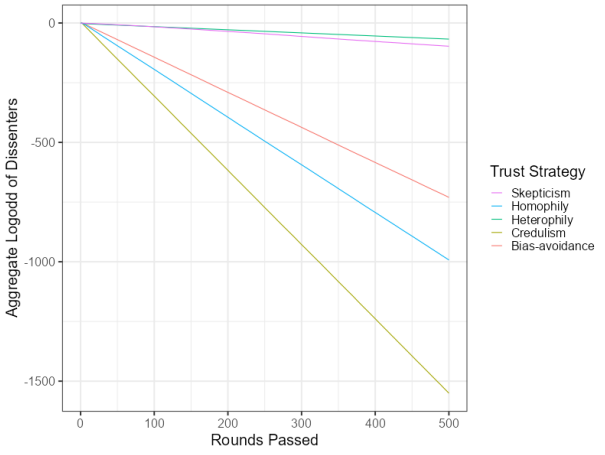

(c) The average aggregate logodd over time




(d) The average and standard deviation of the aggregate logodd after the $1000^{th}$ round of the simulation.

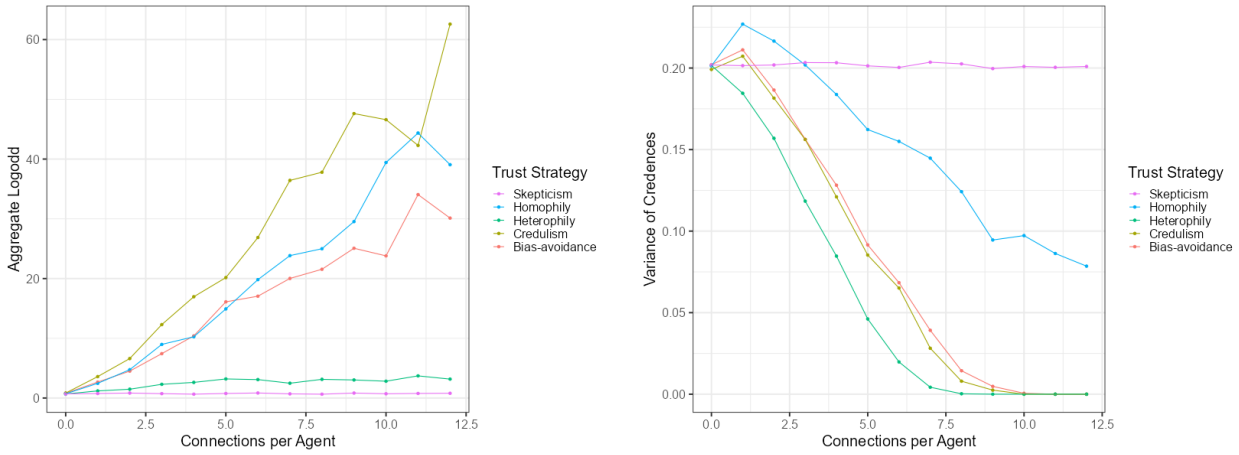



(e) The average aggregate logodd of the dissenters (agents with credence $> .5$) over time




(f) The average aggregate logodd of the dissenters (agents with credence below $\leq .5$) over time

Figure 5.3 The performance of trust strategies on a complete network with biased agents




(a) The average aggregate logodd.

(b) The average variance of credences over time

Figure 5.4 The performance of different trust strategies for various levels of connectivity as of the 50th round of the simulation.

| Examples of communities of agents with extreme (0 or 1) credences | Community 0 | Community 1 | Community 2 | Community 3 | Community 4 | Community 5 | Community 6 | Community 7 | Community 8 | Community 9 | Community 10 | Community 11 | Community 12 | Community 13 | Community 14 | Community 15 | Community 16 | Community 17 | Community 18 | Community 19 | Community 20 | Community 21 | Community 22 | Community 23 | Community 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Agent 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 2 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 3 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 4 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 5 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 6 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Agent 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| Agent 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Agent 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Agent 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| Agent 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Variance of Credences | 0.000 | 0.042 | 0.080 | 0.114 | 0.145 | 0.172 | 0.196 | 0.216 | 0.232 | 0.245 | 0.254 | 0.259 | 0.261 | 0.259 | 0.254 | 0.245 | 0.232 | 0.216 | 0.196 | 0.172 | 0.145 | 0.114 | 0.080 | 0.042 | 0.000 |
| Aggregate Credence | 0.00 | 0.04 | 0.08 | 0.13 | 0.17 | 0.21 | 0.25 | 0.29 | 0.33 | 0.38 | 0.42 | 0.46 | 0.50 | 0.54 | 0.58 | 0.63 | 0.67 | 0.71 | 0.75 | 0.79 | 0.83 | 0.88 | 0.92 | 0.96 | 1.00 |

Figure 5.5 Example communities with agents that each have extreme (0 or 1) credences. In the model, these credences are approached but impossible to reach. But the examples help illustrate what the particular variance values from other figures indicate.

# 6. SHARING-INDUCED ECHO-CHAMBERS & THE MITIGATING EFFECTS OF HETEROPHILY

## Overview

In this chapter, I show that the way in which information is shared can have effects tantamount to manipulating data. I show that even a seemingly altruistic strategy for sharing information, only sharing studies that one believes increases the accuracy of their peers' opinions, can cause echo-chambers which result in entrenchment or group-think depending on other parameters. In addition, I show that agents trusting others to the extent that they disagree with them, heterophily, is able to undermine these echo-chambers.

## 6.1 Introduction

This project does two things: First, it shows the ways in which an agent shares information can have effects tantamount to manipulating that information. Second, it shows how trusting those that one disagrees with, heterophily, can mitigate these effects.

The opinions that members of a community have are largely dictated by the information that those members share amongst each other. Members that are interested in helping their peers increase the accuracy of their opinions have a decision to make about what information they will share. Knowing that fluke studies are possible (those that, while legitimate, erroneously point in the false direction), it might be plausible that the best course of action is for an agent to not share studies that they think are flukes. Agents only sharing studies that they think indicate the truth is called *selective-disclosure* (a.k.a. selective sharing). One might think selective-disclosure is beneficial because it avoids the erroneous fluke studies. And even if it is not beneficial, one might think at the very least it is harmless because all of the studies shared are perfectly legitimate. In this chapter, I show that these intuitions are wrong, and selective sharing can cause issues tantamount to manipulating data. I show that echo-chambers can be caused by selective-disclosure, even when all of it is completely legitimate scientific data (which has not been altered, manipulated or fabricated). These echo-chambers result in challenges nearly identical to ones that can be caused by manipulating data described in the previous chapter: *entrenchment* (where communities are forever split on the issue) and *group-think* (where they unanimously, but arbitrarily agree on an answer).

In addition to this, I show a particular trust strategy, which shows promise in Chapter 4 against echo-chambers, is even more successful here. The strategy is called *heterophily* and involves trusting others in direct proportion with how much they disagree with oneself. The upshot of this project is twofold: i) it shows that the ways that agents share information in their communities —even when it is all completely legitimate— is far from trivial. Instead, simply sharing information in a particular way is tantamount to sharing manipulated infor-

mation. ii) It shows the robustness of heterophily against echo-chambers that have causes other than pure bias, and likewise the value of this strategy for future consideration.

In the next section I explain the background of the project, and afterwards, the model itself. In the subsequent section, I give the results in two parts: First I show the effects of selective-disclosure. Second, I consider how heterophily mitigates these effects.

## 6.2 Background

### 6.2.1 Selective-Disclosure

Selective-disclosure has been studied once before by Weatherall et al., 2020. Their model consists of a scientific community, a propagandist, and a set of non-researchers. The propagandist has access to the data of all of the scientists, while each of the other non-researchers has access to only a subset of the scientists. The propagandist only shares the rare, but existent misleading data of the scientists that indicate the wrong answer. By doing so, the authors show the propagandist is able to hinder the public's ability to learn the truth. They show that the scientists can reach consensus on the correct answer before the non-researchers can. This result helps establish the significance of the sharing strategies used in a community.

However there are a number of limitations to that finding. The authors do not consider performance over time (they simply gauge where the non-researchers are at the time scientists hit a certain threshold), and they also do not consider things like the diversity of the community or the predictability of their opinions. So whether or not the agents become entrenched or fall into group-think is not addressed. It seems to be presumed that the effects of selective sharing are only temporary, or at least, none of their results nor discussion provide reason to think otherwise.

In addition to addressing those concerns, we can also alter their model by considering the strategy in a more neutral setting that does not depend on the existence of a propagandist (who specifically intends to wreak havoc). By considering how the strategy functions when used by the scientists themselves, we get a deeper understanding of the dynamic. It turns

out to be just as dangerous when we do not preemptively assume that only the misleading information is shared. So while their results give a nice first glimpse of the effects of selective sharing (disclosure), much more work is needed to be done to understand the deeper effects. This project sees to that.

This is important to study without presuming misleading information, because this is plausibly the right sort of strategy for achieving the best epistemic consequences. Sharing studies that conflict with the correct opinion (that the new treatment is better) makes one's peers less accurate. So it might seem not only permissible to selectively-disclose, but also maybe imperative if one wants to maximize the epistemic consequences of one's peers.

### 6.2.2 Heterophily

It is a shame that heterophily's fraternal twin, homophily, has hogged so much spotlight in recent literature. Homophily is the common tendency for members of communities to associate with others of like minds. In terms of trust, members trust each other more when they agree with each other more. This strategy has received significant attention for both its positive and negative effects on communities. In Chapter 4, I show that homophily can exacerbate the effects of echo-chambers.

Heterophily, on the other hand, is the tendency for agents to associate with those that one disagrees with. In terms trust, one trusts others the more that they disagree. Another way to think about this is that an agent ignores the information of their peers to the extent that they already have the same opinion. One takes seriously the information they hear to the extent that the agent it comes from has a differing opinion. Whether or not this strategy *is* used in real-world communities, it is important to investigate if it *should* be used. This project lays out the various dynamics of the strategy to help inform this normative question (though it is not the goal to conclusively answer the normative question).

## 6.3 Model

The model concerns a network of agents that share statistically generated evidence to update their opinions concerning a particular problem. The value of this model is that it applies to the broad range of scientific communities that rely on random sampling, and it depends on minimal epistemic assumptions. I will first explain the basics of the model before generalizing it to account for trust.

The problem in question can be thought of as whether or not a particular frequency in nature is above or below a given threshold. One can think of this as doctors determining if a medical treatment's success rate is higher than a known standard, but (as I discuss in Chapter 1), there are many interpretations of the model. All that is important is that the agents used statistically generated evidence to update their opinions. The community consists of $N$ many agents that each produce $n$ new data points every round. The observed frequency of a study is $\frac{k}{n}$ where $k$ denotes the number of successes for the study. The $\frac{k}{n}$ of any given study is likely above or below the true frequency, but overtime the aggregate frequency of the studies reflects it. Agents research every round. This entails the amount of evidence that is produced each round is constant, which ensures that the dynamics that unfold are due purely to the flow of information.

Each agent has a credence indicating the confidence in their opinion that the frequency is above the threshold (that the new treatment of the disease is better). Increasing their accuracy means increasing their credence, since the new treatment is in fact better. I sometimes represent the opinions of agents by applying a logarithmic transformation to their credences. This doesn't require any extra philosophical assumptions and merely provides another useful lens through which to analyze the model. For details, refer to Chapter 1.

Agents share their studies with those they are connected to on a network, according to their sharing strategy. I consider three different sharing strategies:

***Full-disclosure***:

Share all of the studies that you produce.

***Selective-disclosure***:

Share the studies that you produce which you believe reflect the correct opinion.

***Non-disclosure***:

Share none of the studies.

Two of these are simple, indiscriminating strategies. Full-disclosure and non-disclosure treat all studies the same (though the strategies have opposite recommendations for them). Selective-disclosure however shares studies contingently. In the model, this strategy works in the following way: Selective-disclosure agents with credence above 0.5 share their study when it indicates the new treatment is better (when $\frac{k}{n} > .5$). Selective agents with credence at or below 0.5 only share their study when it indicates the treatment is not better (when $\frac{k}{n} \leq .5$).

In Chapter 4, I show how the effects of manipulating information (bias) can be greatly influenced by the amount of connections on a network. The same turns out to be true for selective-disclosure. I demonstrate this with two networks. The cyclic network is one where every agent only has two neighbors, and no two neighbors share a third (assuming the population is $> 3$). The complete network has all agents connected to each other.

Trust can be modelled by having agents discount the evidence they hear from their connections. Each agent $j$ assigns a weight between 0 and 1 for each connection $l$ that they have, denoted $w_{jl}$. Before updating on the evidence they hear, an agent multiplies both $k$ and $n$ by the $w$ they have for that agent. Hence, the level of trust does not change the frequency that the study indicates ($\frac{k}{n} = \frac{w*k}{w*n}$), but it does decrease the level of impact of that study on the agent's opinion. (Note a $k = 60, n = 100$ study impacts an agent's opinion more than a $k = 6, n = 10$ one.) This allows us to redefine the above strategies in terms of $w_{jl}$.

***Credulism***:

$w_{jl} = 1$

***Heterophily***:

$w_{jl} = |c_j - c_l|$

## 6.4 Results

When agents do not share their evidence (non-disclosure), the network structure does not matter, and neither does the trust strategy used. I show in chapter 1 that an agent in such a position is guaranteed to get infinitely confident in the truth as time goes on (based on the evidence that they generate). This helps to serve as a baseline in what follows. The other two strategies, full-disclosure and non-disclosure, are impacted by the level of connectivity in the community as well as the trust strategy used.

I start by analyzing the impacts of the sharing strategies with credulism assumed (for both low and high connectivity), because this is the normal trust strategy used in these models, and it makes the effects of selective-disclosure clear. Afterwards, I consider heterophily as a way to mitigate these effects (for both low and high connectivity). All simulation results can be found in the Additional Materials section.

### 6.4.1 How Selective-Disclosure Causes Echo-Chambers

I first explain how selective-disclosure causes echo-chambers, and the different ways these can manifest. I then explain how the connectivity of the community influences what happens.

#### *6.4.1.1 Echo-chambers & Their Possible Outcomes*

Consider two agents, A and B, connected to each other on a network. There are other agents on the network, but ignore them for the moment. Suppose that A and B both have credences on the same side of 0.5. In any given round, each agent will produce a study that either conforms to their opinion or one that does not. However they will only share these studies when they conform to their opinions. So when agent A observes a study that conforms to their opinion, she shares it with B. B might be connected to other agents as well that share their own studies, and this makes it possible that B will move to the other side

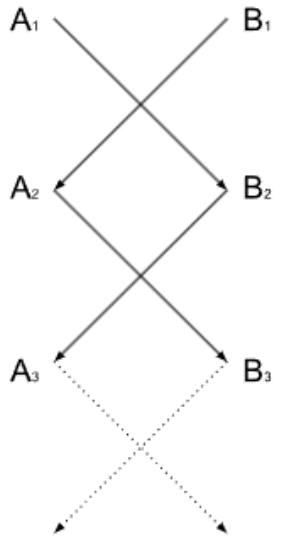



Figure 6.1 Feedback loop between agents A and B over 3 rounds. Arrows indicate the impact of each other's sharing: A's selective-disclosure makes it more likely that B will selectively-disclose consistent information in the following round. Meanwhile B's selective-disclosure makes it more likely that A will do the same.

of .5. But A sharing studies that confirm's B's opinion *increases the likelihood* that on the next round, agent B's credence will still be on the same side of .5. As a result, this makes it more probable that during that subsequent round, B will share a study that also conforms to that view. This would in turn further confirm A's opinion. Likewise, if *B originally shares a study with A*, it will be one that makes it so A is more likely to share a confirming study with B in the following round. Hence, the agents' sharing of only studies that push their credence in a particular direction increases the probability that they will continue to share only studies in that particular direction. This is a *feedback-loop*, and it is the fundamental aspect of how I characterize *echo-chambers*. An illustration of this is given in Figure 6.1. Notice an agent's selective-disclosure does not directly influence their own credence, but it does influence the credence of their neighbor. And so their selective sharing does not affect them in the round that they do it. But it does have an affect on them in the subsequent round when their peer is more likely to 'return the favor'.

Sometimes these feedback loops involve the entire community. When every agent (not just A and B) have similar opinions, and only share evidence in support of that opinion, then they mutually reassure that no agent begins to disagree. This is called *group-think*. Even though an agent's own evidence always is an un-selective sample of the evidence, which might contradict their opinion, the presence of the evidence shared in the community can outweigh this. When it happens, it is permanent. Everyone in the community is determined to believe the opinion they have, and often times it is the wrong one.[1] So agents might believe the wrong thing, even though each of the studies they generate in isolation points to the right answer.[2] Whether or not group-think happens (i.e. everyone collectively decides on an arbitrary answer, whether right or wrong) depends on the amount of connections on the network. I will explain these in just a moment, but it is first important to get a look at the other possible outcomes.

Group-think is not the only possible outcome for the community. Notice that echo-chambers, being defined by feedback-loops, can potentially persist even when the agents are exposed to contrary evidence. If the majority of sources that an agent hears from indicate a particular answer, then that agent will intensify their opinion in that direction even if there is some dissenting information. When a community has disagreeing echo-chambers that persist, this is called *entrenchment*. This can be permanent or subside over time.[3] One way entrenchment can subside is when a larger echo-chamber (in terms of number of agents) has many of its constituents connected to the members of a smaller echo-chamber. Over time, the effect of the larger echo-chamber might win-out against the effects of the smaller one. Regardless of which echo-chamber had which answer, the larger one eventually converts everyone. The result is group-think, since the community is unanimous on an answer that

[1] How often the right versus wrong answer is chosen depends on how difficult the problem is. A more difficult problem makes it easier for the community to unanimously fall to the wrong side of .5. For the remainder of the paper, I will assume that the problem is sufficiently difficult as to make this interesting. The reason for this is reason is because the easier a problem is, the less interesting it is. The dynamics that follow are interesting because they intensify the more difficult that problems get.

[2] This is similar to how Hung and Plott, 2001 describes *information cascades*.

[3] In chapter 4, I include permanence as part of the definition of entrenchment, because there it is not important to distinguish between temporary and permanent forms (since it is all permanent).

could (almost) as easily be wrong as right.

However sometimes the size of echo-chambers are even. In these cases, there are members on the 'edges' of the echo-chambers that are connected to both groups. For example, consider a cyclic network with A and B, and suppose C is connected to B. But suppose that C is also connected to D and they sit on the opposite side of .5 from A and B. This means that B will hear evidence in support of her opinion from A, but evidence against her opinion from B. Which ever of those sources is in support of the right answer will have more opportunity to share studies in support of their view. (There will be more rounds where the study produced actually conforms to the sender's view when the sender is in support of the correct view.) Further, the evidence of B will reflect the truth. This means that B will eventually move to the (or stay on the) correct side. If B was originally wrong, and she moves to the correct side, A becomes the 'edge' agent of the echo-chamber now (assuming that A has other friends that also agree), and the process continues. A is slowly convinced by B's studies and her own. So when echo-chambers are evenly established, they slowly subside in a way that actually reflects the truth.

This sort of slow dissolving of echo-chambers is the only outcome that leaves the agents with plausibly justifiable opinions on the issue. For both group-think and entrenchment, the opinions of the community members are largely arbitrary. The chance that any given agent is increasing their accuracy is a matter of luck. Their opinions are not truly sensitive to original evidence produced, and even when right, they could have easily been wrong. The best a community can hope for is initially even echo-chambers, because these slowly subside in the direction of the truth.

It is also worth mentioning when entrenchment can be permanent, though I will set aside specific exploration of these cases for now. Permanent entrenchment from selective-disclosure turns out to require a particular type of network structure where each agent has an *odd* number of connections. In future work, I plan to investigate the intricacies about how various network structures might allow different manifestations of echo-chambers. For

example, consider agents positioned on the vertices of a cube where the agents on two opposing faces of the cube have opposing opinions. In such a case, the agents hear of their own evidence and one from the the opposing echo-chamber. Meanwhile they hear from two sources that selectively share only in support of their view. In these cases, with sufficiently difficult problems, entrenchment can presumably last forever. But as mentioned, getting into the details is not crucial for this project. The insights for sharing and heterophily are sufficiently established in the other cases (where group-think is permanent and entrenchment is temporary).

#### *6.4.1.2 Network Connectivity*

As mentioned above, whether the community enters into group-think or entrenchment depends significantly on the amount of connectivity on the network. To see this, it is helpful to recapitulate the results for credulist, fully-disclosive communities covered in Chapter 1. When such agents have more connections, they have access to more information. When agents are part of complete networks, they all have exactly the same pools of evidence. I show in that chapter that this leads to a less diverse and less predictable community. While the community is guaranteed to find the truth, the path by which they get there is less clear. There is essentially no room for difference of opinions, and the aggregate opinion is more sensitive to misleading series of studies. On a cyclic network, the agents consume much less information each round. The diversity of the opinions reflects the diversity of the evidence itself, and so the aggregate opinion is more predictable.

While increasing connections makes a community less predictable, it is not without a feature like selective-disclosure (or bias covered in Chapter 4) that the community is at risk of not finding the truth. And the potential outcome for the community fits with the results of chapter 1: Increasing connections decreases the chance for any diversity of opinion. So when agents are selective sharers on a complete network, they are doomed to engage in group think. Consider Figure 6.3 which depicts multiple metrics of the community. In particular, notice that the variance of credences (Fig. 6.3b) of the selective-disclosure vanishes even

more quickly than that of the fully-disclosure community. Meanwhile the aggregate opinion of the community is highly unpredictable, as shown by the standard deviation (Fig. 6.3d) of the logodds (which are a logarithmic representation of credences). This means that aggregate opinions shown in Figures 6.3c and 6.3a do not reflect the aggregate opinions of any particular run, but instead averaging communities that have very extreme opinions.[4] These extreme opinions are shown in the aggregate logodds when everyone believes the new treatment is better (*assenter*) versus when everyone does not (*dissenter*) as seen in Figures 6.3e and 6.3f.

On the other hand decreasing connectivity means agents have less access to each other's data, and there is more potential for epistemic diversity. Hence echo-chambers can arise like those shown in Figure 6.2. First notice that the variance of the selective communities not only does not vanish like it does with high connectivity, it is higher than that of the other strategies (Figure 6.2b). This shows that the agents are engaging in contrasting echo-chambers, not only ones containing the entire community. This means that the aggregate credence (Figure 6.2a) and logodds (Figure 6.2c) are much closer to the actual values of particular runs. In these cases, the assenters (6.2e) and dissenters (6.2f) can occur in the same simulation. It is also worth remembering at this point that the entrenchment here is not permanent. So that is why the diversity of opinions does slowly diminish. In these cases, the dissenters that continue to go down forever represent simulations that resulted in group-think.

These results show how selective-disclosure causes echo-chambers, and how these echo-chambers can result in two seemingly very different outcomes. When there is more connectivity, it is hard for there to be any diversity of opinions, and the result is group-think. When communication is more sparse, the potential for disagreeing echo-chambers to persist (entrenchment) increases. In both cases, the selective communities do not accomplish achieving the epistemic ends mentioned above that might make it seem like a plausible strategy: the strategy is not increasing the chance that others in one's community have more accurate

[4]Note that credences and logodds are two representations of the same quantity. Logodds are credences after a logarithmic transform.

credences.

The above shows how echo-chambers that persist do so because there is an over abundance of one type of evidence (i.e. evidence that supports one direction, whichever direction that is). The problem here is that the studies can no longer be thought of as neutral — even thoug h there is nothing about the studies themselves that makes them non-neutral. The 'non-neutrality' comes in at the point where agents select for only a certain range of studies to be shared. They assume that their opinion is one that is an undistorted reflection of the truth, and based on that assumption make the plausible move to only share studies that would increase the accuracy of their peers. At the heart of the issue is that an agent mistake their own opinion as neutral and well-informed and thereby cause others in their community to make the same mistake. The result is a community that unknowingly engaged in a socially extended version of circular reasoning.

Still, whether or not people *should* selectively share, they still do it. And so it is important to find strategies to mitigate this sort of sharing.

### 6.4.2 How Heterophily Undermines the Effects of Selective-Disclosure

#### *6.4.2.1 General Explanation*

It is worthwhile to first note that one trust strategy to go with is simply skepticism. This makes the sharing strategies of agents a non-issue, and so each agent is guaranteed to find the truth. But this does not settle the issue. It is likely an overreaction to eliminate communication (or the effects of it) entirely to avoid problems. This strategy should be remembered for contingency purposes, but scientific communities are likely better off with a more nuanced strategy that allows them to navigate what is being shared. (It is also worth pointing out that skepticism sounds implausible to achieve fully.)

Findings from Chapters 3 and 4 suggest that a strategy for (dis)trusting others might be an effective way to mitigate the impacts of selective-disclosure. As discussed above, heterophily involves trusting others to the extent that one disagrees with them: The closer

that agents are in credence, the less weight they put on the evidence they hear from each other. Meanwhile, the further they are in credence, the more weight will be put on the evidence.

Consider the example above with A and B. When A and B have similar opinions, the risk with selective-disclosure is that they form a feedback loop by mutually encouraging each other to continue to share studies that support their opinion. But with heterophily, the more their opinions grow together, the less they are affected by each other's evidence. This undermines the potential for an echo-chamber to form between A and B.

Now, it is still possible (if only temporarily) that an echo-chamber arises. Because there is still some trust between A and B, and for a limited number of runs[5] this can facilitate their increasing of each other's opinions in the same direction. For communities of higher connectivity, these echo-chambers can persist for longer, because even while diminished there is still more trust and information flowing than in less connected communities.

But consider the impacts of B being connected to C, an agent form outside its echo-chamber. To the extent that B and C have differing opinions (which is intensified if one or both of them are in opposing echo-chambers), they trust each other's evidence more. This means that C's evidence actively pulls B out of its echo-chamber more strongly the deeper that B goes. So not only does heterophily curtail the formation of echo-chambers, it actively works to eliminate them.

#### *6.4.2.2 Demonstrated on the Cyclic Network*

Figure 6.4 shows these results for the cyclic network. Most notable is the variance of the community. While the variance of the selective community does still initially increase, it does not reach the same level it does when credulism is in play. (With credulism, the peak is near .16, with heterophily it is closer to .09.) Further, the rate at which that variance disappears is much quicker for the heterophilic community. Without the significant effects

[5]Technically, there is no final cutoff for when there is no trust and likewise no effect. That is because the change in trust is continuous. But there is an intuitive sense in which we can say the effect becomes insignificant.

of echo-chambers, the selective community performs much more similarly to the full- and non-disclosure communities. Notice how the aggregate opinions (Figures 6.4a and 6.4c) of the three sharing strategies are quite close together (compared to when credulism is used), and similarly, the predictabilities of the aggregate logodds (6.4d) are closer. This is because in the long run, every agent approaches the same credence (1). Hence heterophily demands that their trust in each other eventually all go to zero.

So what is the difference in performance between heterophily, which effectively turns into skepticism, and skepticism from the get-go? The difference concerns how the early rounds, where there is trust, effect the performance in the long term. Notice that selective-disclosure achieves higher aggregate opinions quicker than non-disclosure (which is effectively skepticism) specifically due to the early rounds where agents did trust each other to some extent. So there is potentially value in their being some trust between agents. Heterophily shows how one can capture some of that value even in the midst of selective-disclosure. Further, this suggests the need to explore even more nuanced strategies that undermine echo-chambers similar to how heterophily does, but allow agents non-negligible levels of trust in the long term. This is the sort of strategy I plan to search for in future work.

#### *6.4.2.3 Demonstrated on the Complete Network*

Figure 6.5 shows the effects of heterophily on a complete network. While heterophily collapses the difference between disclosure strategies on the cyclic network, it does not do so on the complete network. The difference comes down to the fact that the increased connectivity can make stronger initial echo-chambers than on a cyclic network due to there being agents having more opportunity to engage in feedback loops. But even more significant to the effects is the absence of dissenting members. In the terms of the example above, there is no C to actively bring B out of their echo-chamber (since C would be part of the same echo-chamber). This is why in 6.5b, one can see the variance (almost) immediately disappears.

What this means is that the entire community, at least in the initial rounds, engages in group-think, where they unanimously move to one side of .5. No one is left on the opposite

side, because there would necessarily be trust between them as a result. And that trust forces anyone with a minority view to make their way back to the majority side. (Note a minority member would get evidence from every member of the majority, because it is a complete network, but little evidence from fellow minority members -if there were any- due to heterophily.) So what we have is a community that quickly makes its way to one side of .5, but in doing so, achieves a more similar credence. This means the agents stop trusting each other, and they do not accelerate into their unanimously chosen position like they do with credulism. Instead, since everyone begins to distrust each other, they only update on their own evidence. At this point, the behavior of the community depends on which side of .5 they happened to end up on.

Suppose the community is above .5 (which is the side that reflects the truth). That means each agent in the community will individually produce evidence which increases their credence. Since each agent is expected to move independently in the same direction, they tend to keep similar credences, and trust (nor each other's evidence) comes into play. If there does happen to be agents that get fluke studies pointing them in the opposite direction, then trust kicks in and the other members 'save' their fledgling comrade. The community is never expected to switch to the other side of .5, because doing so would require the independent generation of fluke studies by the majority of individuals in the community. This is why the assenters in Figure 6.5e perpetually increase their opinions. In these cases, the community is unanimously, but essentially independently[6] approaches credence 1. Hence the community in these cases approaches 1, but probably not in a way that can be characterized a group-think: beyond the initial rounds (which pushed everyone to one side), there are no feedback-loops between agents. Before discussing whether or not this is legitimate group-think, or if it reflects behavior that is somehow rational for the community, it is important to see how the other case is handled.

[6]That the possibility of dissent is met with trust is like bumpers in a bowling alley. Agents typically make their way to the goal on their own and in normal cases will not have noticed that the bumpers (possibility of trust) was even there. But when an agent does need it, when they skew from their peers, it ensures they bounce back in the direction they need to go.

As mentioned above, even though echo-chambers are quickly dissolved by heterophily, heterophily on a complete network still forces everyone to one side on .5, and sometimes that is the wrong side (below .5). But notice what happens when this is the case. Just like if they were above .5, the trust between agents pushes them all to the same credence. Like before, trust essentially disappears and only reappears to quickly bring disagreeing agents back in line. So when everyone has a similar credence, they depend on their own evidence to update. So far, this is equivalent to when the community happened to all land above .5. The difference comes in with the independent evidence that each agent generates. When in virtual isolation (no trust due to similar credences), the evidence of each agent will *not* push them further toward the wrong answer. That is because their (unselected) information reflects the truth, and tells them to break from the community and increase their credence. This is why in Figure 6.5f, the dissenters to not constantly dig in their heels. (Note the difference in scales of Figure 6.5f versus Figure 6.3f.) What happens is the community plateaus not far from a credence of .5 (how far is determined by the difficulty of the problem, but specific values do no matter here). The community cannot go lower, because doing so would require the individual evidence to make that happen. Heterophily undermines the feedback loops that allow for the community to dig in their heels by distorting the distributions of the evidence. In fact the individually produced studies of the community have the potential to move the community independently and simultaneously up past .5. Since the communities never stray too far from .5 in the first place, this increases the chances that the community will be able to pass the threshold of .5 and continue increasing their accuracy. They cannot move too quickly, because while below .5 if any agents move faster than others, they will be on the receiving end of selected studies that pull them back down. But so long as agents each receive as informative information as the next, they can all make the same marginal gains toward the truth without peer pressure.

This shows an interesting feature of heterophily: even though group-think can initially occur in either the right or wrong direction, the feedback-loops quickly stop. In what follows,

there is an asymmetry to the behavior of the community depending on whether or not they found the right answer. This means that one can consider the opinions of the correct communities safe and sensitive to evidence in a way that could not be attributed to the group-think credulist communities (even when they were right). The performance of the heterophilic community depends on (is sensitive to) which answer is right. Likewise, opinions that increase rapidly are safe, because they could not be achieved with the wrong answer.

## 6.5 Conclusion

There are four main upshots of this project.

First, sharing even completely legitimate information in a particular way can have the same effects of manipulating data. A conclusion one might make from this is that parsing up information to share based on the studies themselves is not the right level of granularity to make decisions about sharing information. Instead of making decisions study by study, it might be the case that we should make a courser grained decision at the level of *problems*. That is, if one is going to share one study concerning a specific problem, then they ought to share all of their studies concerning that problem.

However there are other alternatives to selective-disclosure than simply full- or non-disclosure. One example is a more nuanced strategy, such as one where studies are shared in accordance with the scientist's estimation of the frequency, and not the credence concerning whether or not that frequency is above a certain threshold. Selective disclosure is the right place to start, because it has (had) a plausible claim to being the right consequentialist strategy. In future work, I plan to investigate more nuanced strategies that might also be consequentially or otherwise motivated.

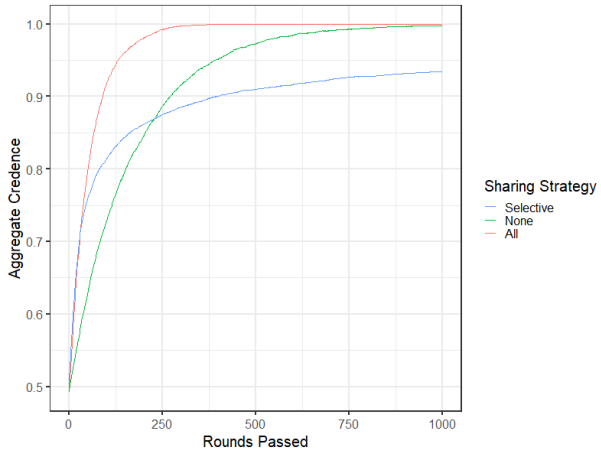
Second, aside from showing this particular issue, this demonstrates the more general lesson that disclosure strategies are far from trivial. This is important because i) how unknown this seems to be, and ii) how much impact sharing can have. Most people would likely not see an issue in refusing to share studies they think are flukes (while still sharing the rest that they see). And because there is such a strong potential impact, it makes this feature

particularly insidious.

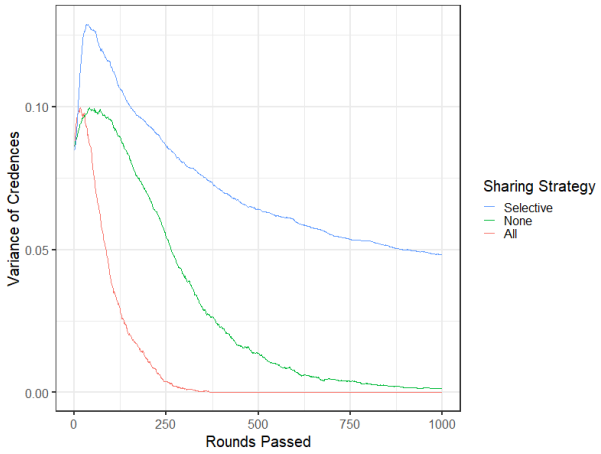
Third, this project yields a novel way to evaluate trust strategies. Aside from the variety of metrics that are used, we were able to discover an interesting feature of heterophily in its assymmetry of dealing with right versus wrong answers. So not only should trust strategies be evaluated with regard to higher or lower values on specific metrics, but also with regard to whether they fundamentally treat certain answers differently.

Fourth, this project further demonstrates the capabilities of heterophily which have already been extolled in Chapters 3 and 4. This shows that the performance of heterophily is not an accident. Creating correlation between those that disagree and eliminating between those that agree is proving to be one of the most interesting (albeit simple) way of navigating scientific variance.

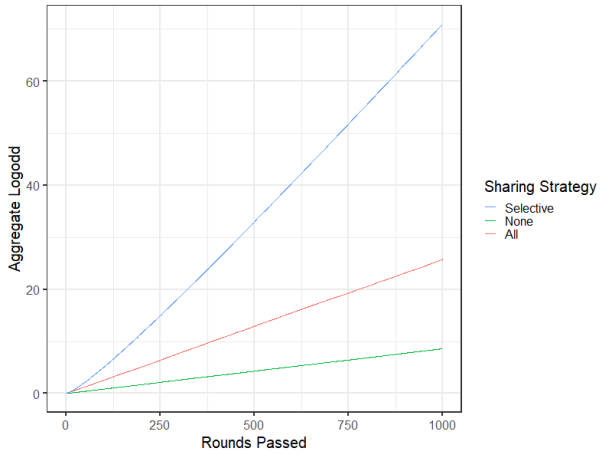
## 6.6 Additional Material



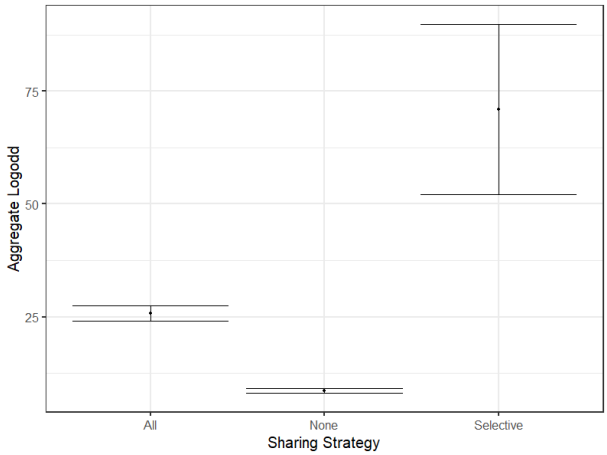

(a) The average aggregate credence over time



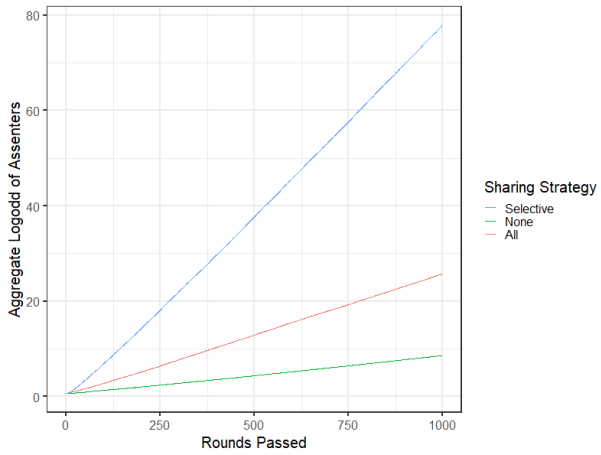

(b) The average variance of credences over time



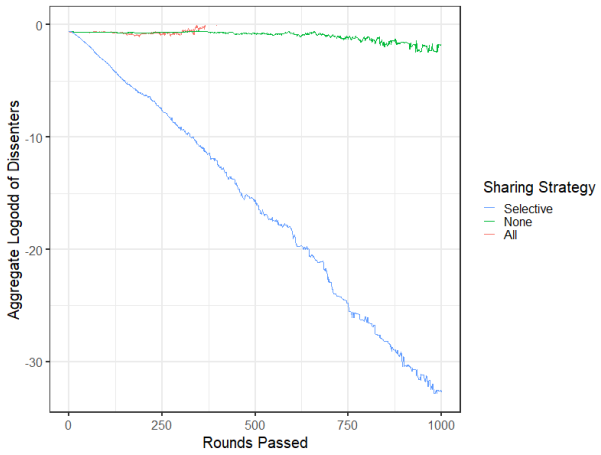

(c) The average aggregate logodd over time




(d) The average and standard deviation of the aggregate logodd after the $1000^{th}$ round of the simulation.

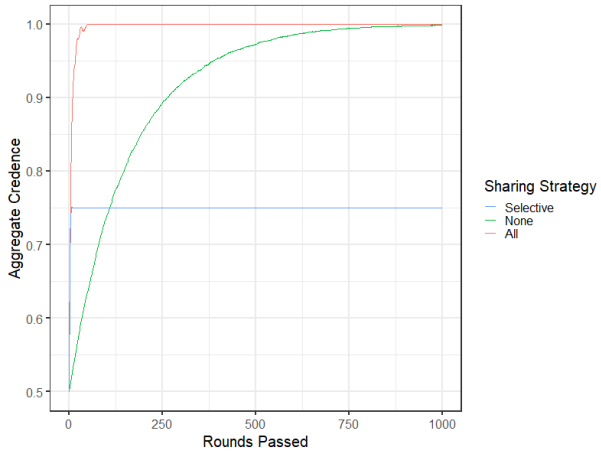



(e) The average aggregate logodd of the dissenters (agents with credence $> .5$) over time

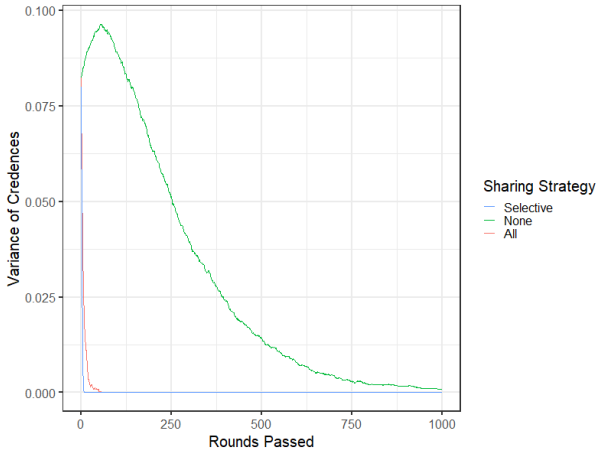



(f) The average aggregate logodd of the dissenters (agents with credence below $\leq .5$) over time

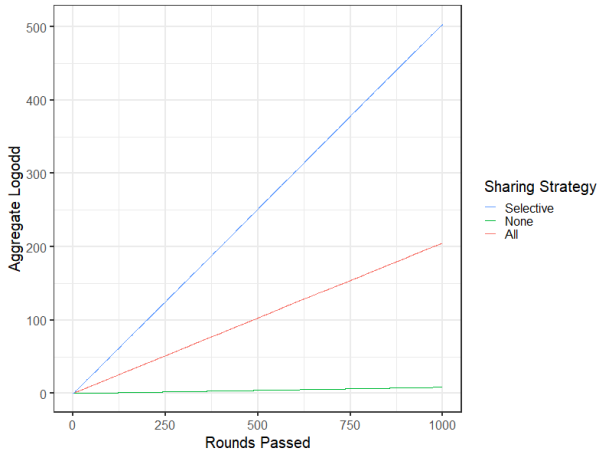
Figure 6.2 The performance of sharing strategies on a cyclic network with credulist agents



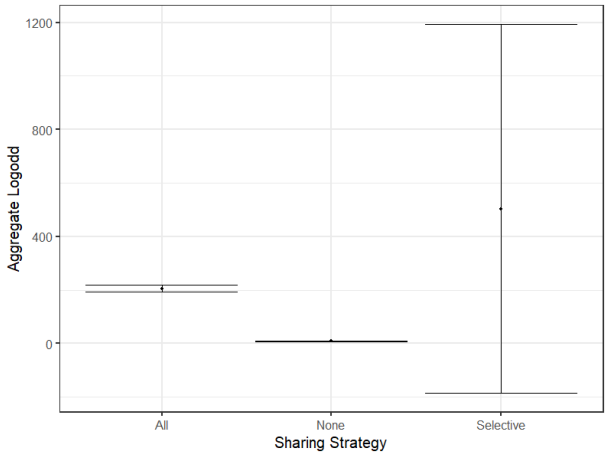

(a) The average aggregate credence over time



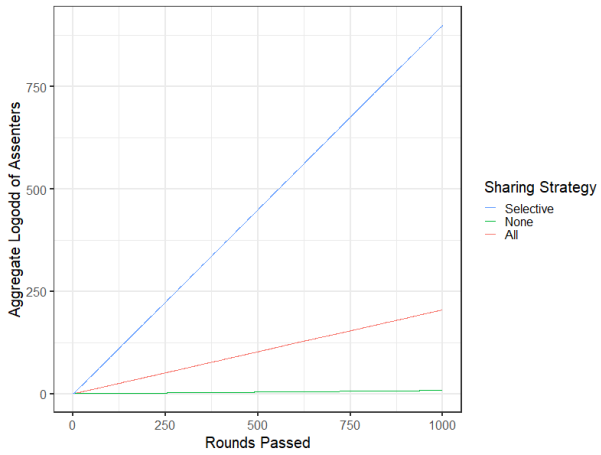

(b) The average variance of credences over time



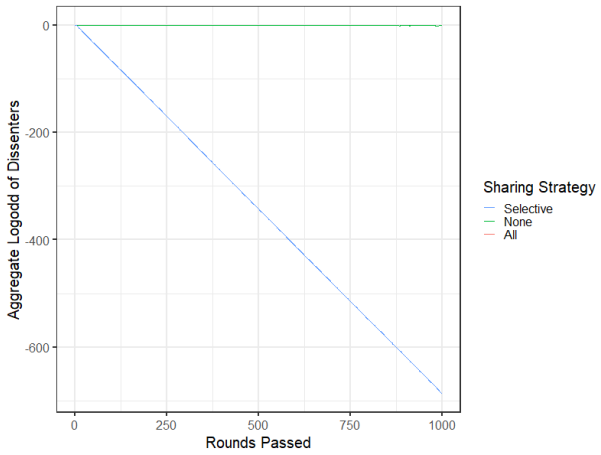

(c) The average aggregate logodd over time




(d) The average and standard deviation of the aggregate logodd after the $1000^{th}$ round of the simulation.

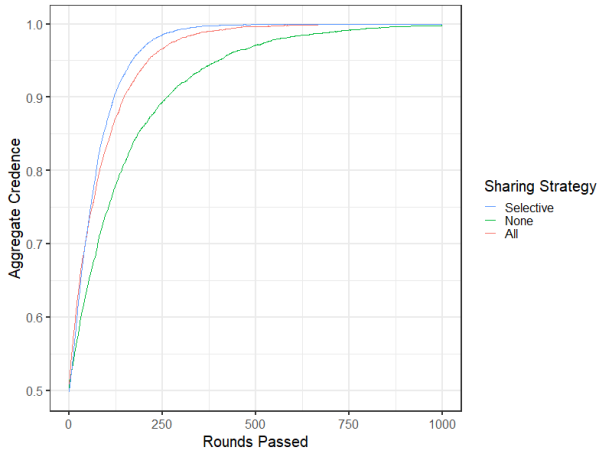



(e) The average aggregate logodd of the dissenters (agents with credence $> .5$) over time

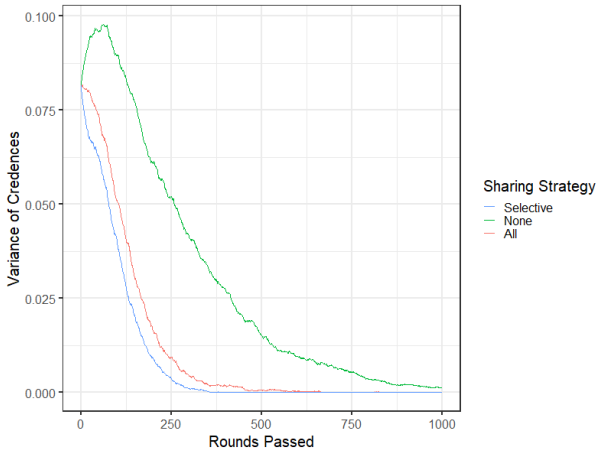



(f) The average aggregate logodd of the dissenters (agents with credence below $\leq .5$) over time

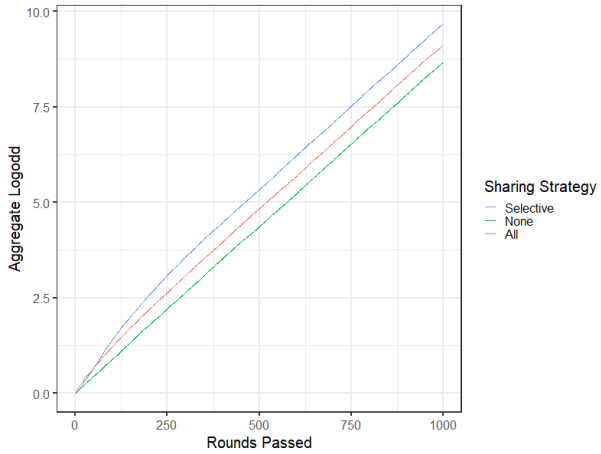
Figure 6.3 The performance of sharing strategies on a complete network with credulist agents



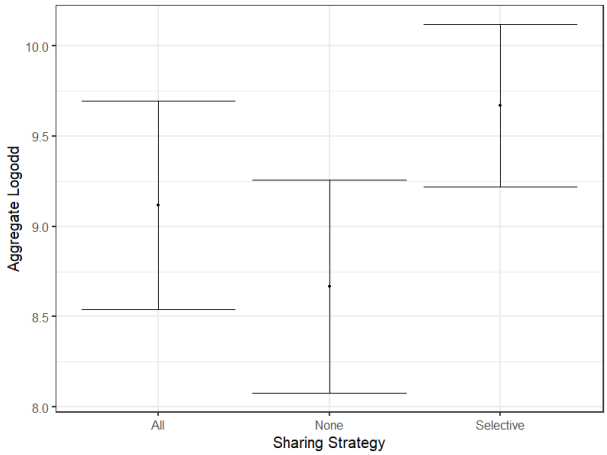

(a) The average aggregate credence over time



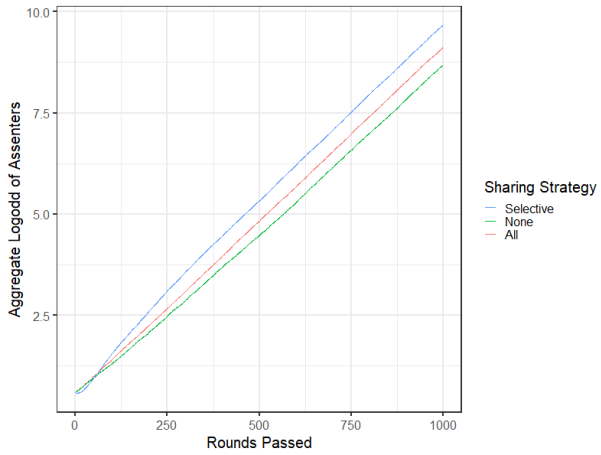

(b) The average variance of credences over time



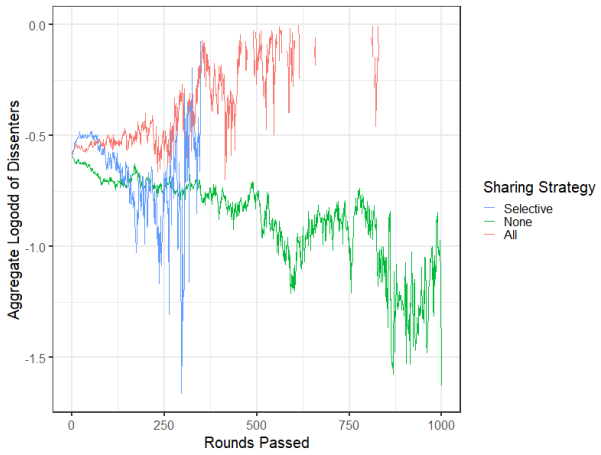

(c) The average aggregate logodd over time




(d) The average and standard deviation of the aggregate logodd after the $1000^{th}$ round of the simulation.

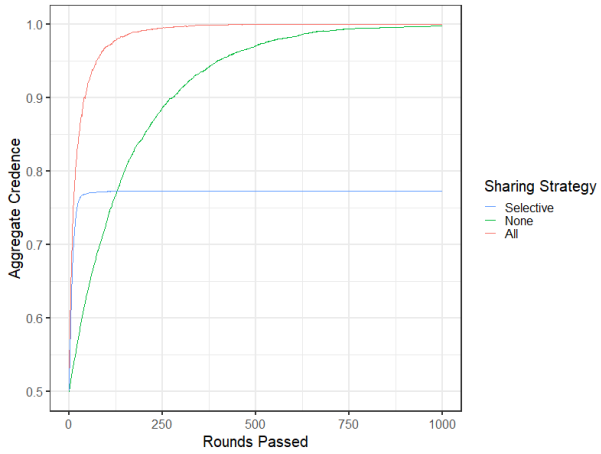



(e) The average aggregate logodd of the dissenters (agents with credence $> .5$) over time

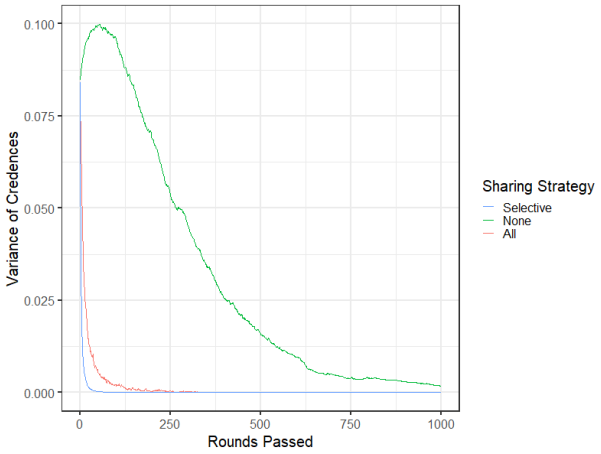



(f) The average aggregate logodd of the dissenters (agents with credence below $\leq .5$) over time

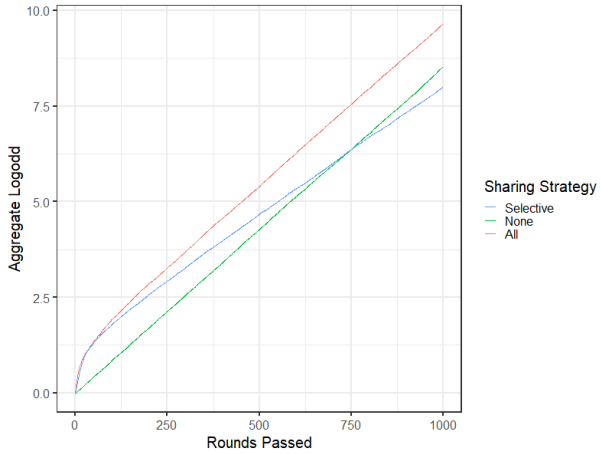
Figure 6.4 The performance of sharing strategies on a cyclic network with heterophilic agents



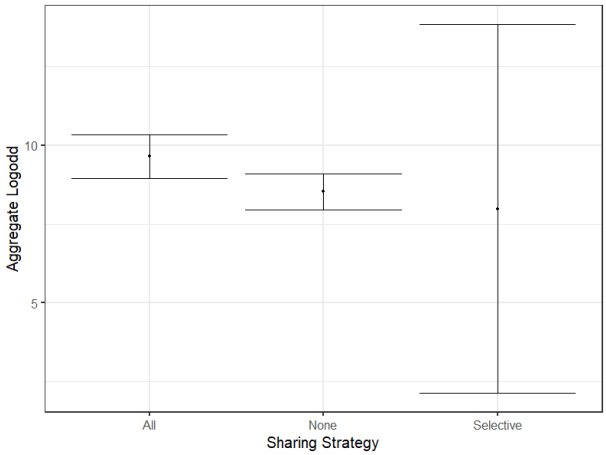

(a) The average aggregate credence over time



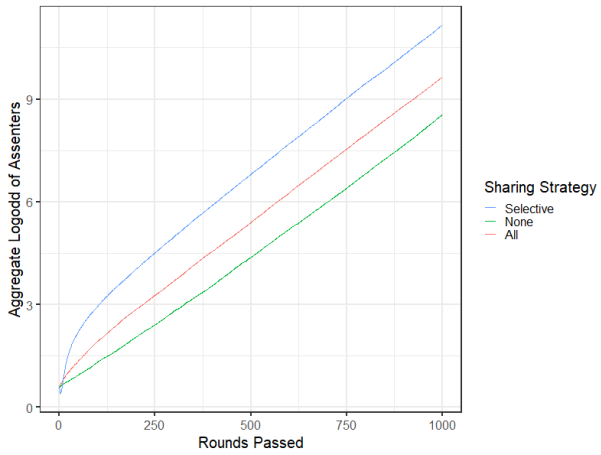

(b) The average variance of credences over time



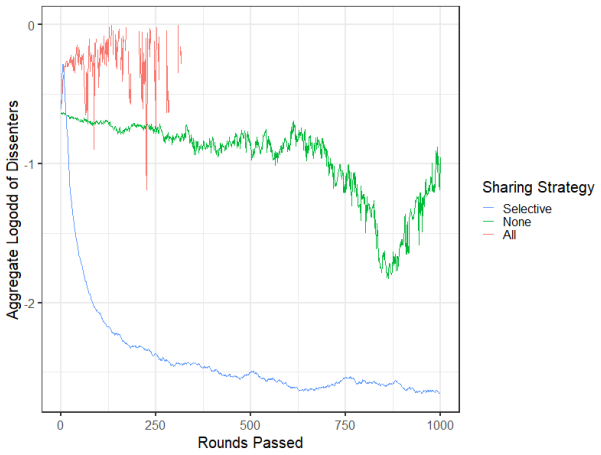

(c) The average aggregate logodd over time




(d) The average and standard deviation of the aggregate logodd after the $1000^{th}$ round of the simulation.




(e) The average aggregate logodd of the dissenters (agents with credence $> .5$) over time




(f) The average aggregate logodd of the dissenters (agents with credence below $\leq .5$) over time

Figure 6.5 The performance of sharing strategies on a complete network with heterophilic agents

# REFERENCES


Bala, V., & Goyal, S. (1998). Learning from Neighbours. *Review of Economic Studies*, *65*(3), 595–621. https://doi.org/10.1111/1467-937X.00059

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press. https://doi.org/10.1017/CBO9780511616037

Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures [Publisher: Routledge _eprint: https://doi.org/10.1080/0022250X.2016.1147443]. *The Journal of Mathematical Sociology*, *40*(2), 80–111. https://doi.org/10.1080/0022250X.2016.1147443

Bright, L. K. (2021). Why Do Scientists Lie? [Publisher: Cambridge University Press]. *Royal Institute of Philosophy Supplements*, *89*, 117–129. https://doi.org/10.1017/S1358246121000102

Easwaran, K., Fenton-Glynn, L., & Hitchcock, C. (2016). Updating on the Credences of Others: Disagreement, Agreement, and Synergy. *16*(11), 39.

Fazelpour, S., & Steel, D. (2022). Diversity, Trust and Conformity: A Simulation Study. *Philosophy of Science*.

Fricker, E. (1994). Against Gullibility. In A. Chakrabarti & B. K. Matilal (Eds.), *Knowing from Words*. Kluwer Academic Publishers.

Golub, B., & Jackson, M. O. (2012). How Homophily Affects the Speed of Learning and Best-Response Dynamics. *The Quartarly Journal of Economics*, *127*(3), 1287–1338. https://doi.org/10.1093/qje/qjs021

Grim, P., Singer, D. J., Bramson, A., Holman, B., McGeehan, S., & Berger, W. J. (2019). Diversity, Ability, and Expertise in Epistemic Communities [Publisher: The University of Chicago Press]. *Philosophy of Science*, *86*(1), 98–123. https://doi.org/10.1086/701070

Haghtalab, N., Jackson, M. O., & Procaccia, A. (2020). *Belief Polarization in a Complex World: A Learning Theory Perspective* (SSRN Scholarly Paper No. ID 3606003). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3606003

Heesen, R. (2017a). Communism and the incentive to share in science. *Philosophy of Science*, *84*(4), 698–716. https://doi.org/10.1086/693875

Heesen, R. (2017b). Academic Superstars: Competent or Lucky? *Synthese: an international journal for epistemology, methodology and philosophy of science*, *194*(11), 4499–4518. http://dx.doi.org/10.1007/S11229-016-1146-5

Heesen, R. (2018). When journal editors play favorites. *Philosophical Studies*, *175*(4), 831–858. https://doi.org/10.1007/s11098-017-0895-4

Heesen, R., & Bright, L. K. (2021). Is Peer Review a Good Idea? [Publisher: The University of Chicago Press]. *The British Journal for the Philosophy of Science*, *72*(3), 635–663. https://doi.org/10.1093/bjps/axz029

Holman, B., & Bruner, J. P. (2015). The Problem of Intransigently Biased Agents. *Philosophy of Science*, *82*(5), 956–968. https://doi.org/10.1086/683344

Holman, B., & Geislar, S. (2018). Sex Drugs and Corporate Ventriloquism: How to Evaluate Science Policies Intended to Manage Industry-Funded Bias [Publisher: The University of Chicago Press]. *Philosophy of Science*, *85*(5), 869–881. https://doi.org/10.1086/699713

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

Hung, A. A., & Plott, C. R. (2001). Information Cascades: Replication and an Extension to Majority Rule and Conformity-Rewarding Institutions. *American Economic Review*, *91*(5), 1508–1520. https://doi.org/10.1257/aer.91.5.1508

Kitcher, P. (1990). The Division of Cognitive Labor [Publisher: Journal of Philosophy, Inc.]. *The Journal of Philosophy*, *87*(1), 5–22. https://doi.org/10.2307/2026796

Kuhn, T. S. (1970). *The structure of scientific revolutions* ([2d ed., enl). University of Chicago Press.

Kuhn, T. S. (1977). Objectivity, Value Judgment, and Theory Choice.

Kummerfeld, E., & Zollman, K. J. S. (2016). Conservatism and the Scientific State of Nature. *The British Journal for the Philosophy of Science*, *67*(4), 1057–1076. https://doi.org/10.1093/bjps/axv013

Lackey, J. (2007). Norms of Assertion [Publisher: Wiley]. *Noûs*, *41*(4), 594–626. https://www.jstor.org/stable/4494552

Lackey, J. (2008). *Learning from Words: Testimony as a Source of Knowledge* [Publication Title: Learning from Words]. Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199219162.001.0001/acprof-9780199219162

List, C., & Goodin, R. E. (2001). Epistemic Democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, *9*(3), 277–306. https://doi.org/10.1111/1467-9760.00128

Lord, C., Ross, L., & Lepper, M. (1979). Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Mayo-Wilson, C. (2014). Reliability of testimonial norms in scientific communities. *Synthese*, *191*(1), 55–78. https://doi.org/10.1007/s11229-013-0320-2

Mayo-Wilson, C., Zollman, K. J. S., & Danks, D. (2011). The Independence Thesis: When Individual and Social Epistemology Diverge* [Publisher: [The University of Chicago Press, Philosophy of Science Association]]. *Philosophy of Science*, *78*(4), 653–677. https://doi.org/10.1086/661777

Mohseni, A., & Williams, C. R. (2021). Truth and Conformity on Networks. *Erkenntnis*, *86*(6), 1509–1530. https://doi.org/10.1007/s10670-019-00167-6

Nguyen, C. T. (2020). ECHO CHAMBERS AND EPISTEMIC BUBBLES. *Episteme*, *17*(2), 141–161. https://doi.org/10.1017/epi.2018.32

O'Connor, C., & Gabriel, N. (2022). *Can Confirmation Bias Improve Group Learning?* (preprint). MetaArXiv. https://doi.org/10.31222/osf.io/dzych

O'Connor, C., & Weatherall, J. O. (2018). Scientific polarization. *European Journal for Philosophy of Science*, *8*(3), 855–875. https://doi.org/10.1007/s13194-018-0213-9

O'Connor, C., & Wu, J. (2021). *How Should We Promote Transient Diversity in Science?* (preprint). MetaArXiv. https://doi.org/10.31222/osf.io/w3xc5

Rosenstock, S., Bruner, J., & O'Connor, C. (2017). In Epistemic Networks, Is Less Really More? *Philosophy of Science*, *84*(2), 234–252. https://doi.org/10.1086/690717

Singer, D. J. (2019). Diversity, Not Randomness, Trumps Ability. *Philosophy of Science*, *86*(1), 178–191. https://doi.org/10.1086/701074

Singer, D. J., Bramson, A., Grim, P., Holman, B., Jung, J., Kovaka, K., Ranginani, A., & Berger, W. J. (2019). Rational social and political polarization. *Philosophical Studies*, *176*(9), 2243–2267. https://doi.org/10.1007/s11098-018-1124-5

Soysal, Z. (2019). Truth in Journalism. In *Journalism and Truth in an Age of Social Media*. Oxford University Press. https://doi.org/10.1093/oso/9780190900250.003.0008

Strevens, M. (2003). The Role of the Priority Rule in Science: *Journal of Philosophy*, *100*(2), 55–79. https://doi.org/10.5840/jphil2003100224

Strevens, M. (2020). *The Knowledge Machine: How Irrationality Created Modern Science*. New York: Liveright Publishing Corporation.

Thompson, A. (2014). Does Diversity Trump Ability? An Example of the Misuse of Mathematics in the Social Sciences. *Notices of the AMS*, *61*(9), 1024–1030. https://doi.org/http://dx.doi.org/10.1090/noti1163

Weatherall, J. O., & O'Connor, C. (2020). Endogenous epistemic factionalization. *Synthese*. https://doi.org/10.1007/s11229-020-02675-3

Weatherall, J. O., O'Connor, C., & Bruner, J. P. (2020). How to Beat Science and Influence People: Policymakers and Propaganda in Epistemic Networks [Publisher: The University of Chicago Press]. *The British Journal for the Philosophy of Science*, *71*(4), 1157–1186. https://doi.org/10.1093/bjps/axy062

Weisberg, M., & Muldoon, R. (2009). Epistemic Landscapes and the Division of Cognitive Labor*. *Philosophy of Science*, *76*(2), 225–252. https://doi.org/10.1086/644786

Wu, J. (2022). Epistemic advantage on the margin: A Network Standpoint Epistemology. *Philosophy and Phenomenological Research*, phpr.12895. https://doi.org/10.1111/phpr.12895

Zollman, K. (2009). Optimal Publishing Strategies. *Episteme*, *6*, 185–199. https://doi.org/10.3366/E174236000900063X

Zollman, K. J. S. (2007). The Communication Structure of Epistemic Communities. *Philosophy of Science*, *74*(5), 574–587. https://doi.org/10.1086/525605

Zollman, K. J. S. (2010). The Epistemic Benefit of Transient Diversity. *Erkenntnis*, *72*(1), 17–35. https://doi.org/10.1007/s10670-009-9194-6

Zollman, K. J. S. (2015). Modeling the social consequences of testimonial norms. *Philosophical Studies*, *172*(9), 2371–2383. https://doi.org/10.1007/s11098-014-0416-7

Zollman, K. J. (2012). Social network structure and the achievement of consensus. *Politics, Philosophy & Economics*, *11*(1), 26–44. https://doi.org/10.1177/1470594X11416766