Reliability of Testimonial Norms in Scientific Communities

Conor Mayo-Wilson

July 19, 2013

Abstract

Several current debates in the epistemology of testimony are implicitly motivated by concerns about the reliability of rules for changing one's beliefs in light of others' claims. Call such rules testimonial norms (TNs). To date, epistemologists have neither (i) characterized those features of communities that influence the reliability of TNs, nor (ii) evaluated the reliability of TNs as those features vary. These are the aims of this paper. I focus on *scientific* communities, where the transmission of highly specialized information is both ubiquitous and critically important. Employing a formal model of scientific inquiry, I argue that miscommunication and the "communicative structure" of science strongly influence the reliability of TNs, where reliability is made precise in three ways.

Introduction

Most of our scientific knowledge is based upon others' testimony. For instance, I have never verified Coulomb's law, but I know that electrostatic force obeys an inverse square law. Why? My high school physics teacher told me so. On first glance, explaining how I know such facts seems easy. Scientific facts, like Coulomb's law, are first confirmed by experts in an experiment. Experts then disseminate their findings by word of mouth or through journals. Those findings are, in turn, summarized in survey articles for other academics, in textbooks for students, and in popular media. In this way, experts' findings, like Coulomb's law, are transmitted from one person to another. Each part of this explanation, however, raises serious questions. I will mention two.

First, scientists often disagree. How can non-specialists justifiably accept a hypothesis when there are experts who disagree? Recently, epistemologists have developed several procedures for evaluating experts and deciding whom to trust. Some argue that, absent other information, non-specialists should "go by the numbers" and adopt the opinion of the majority of experts in a field. Others argue that there are criteria by which non-experts can determine which experts are most reliable.[1]

Second, much of our scientific knowledge is acquired from non-experts. Journalists often have no scientific training, and yet both academics and lay audiences often learn of scientific advances via newspapers and magazines. Many secondary teachers and college professors are not experts in the field in which they provide instruction. And so on. Because non-experts may be (i) unreliable, (ii) dishonest, and/or (iii) prone to miscommunication, there are serious questions concerning who one should trust.

Recently, epistemologists have defended two theses, called reductionism and non-reductionism respectively, that attempt to characterize when one is justified in believing others. Recognizing the implausibility of always verifying others' claims, non-reductionists argue that one may justifiably trust a speaker in the absence of evidence of dishonesty or unreliability. In contrast, recognizing the frequency of dishonesty and/or unreliability, reductionists argue that one needs *positive reasons* to trust others, where such positive reasons might include evidence for the speakers' honesty and/or expertise. And there are philosophers who adopt intermediate positions.[2]

The above debates are partly motivated by concerns about the reliability of rules for changing one's beliefs in light of others' claims. Call such rules **testimonial norms** (TNs). The rule "believe others in the absence of conflicting information" is one TN, and "only believe those you know to be reliable" is another. Arguably, the reductionism debate is, in part, motivated by the observation that the former TN is reliable in certain contexts but not in others, and the latter prohibits one from learning when speakers' are reliable but cannot known to be so. Similar remarks apply to debates about expert testimony. It is surprising, therefore, that epistemologists have made little effort (i) to characterize those contextual features that influence the reliability of TNs, or (ii) to evaluate the reliability of TNs as those features

[1] The Lehrer-Wagner model entails that, *ceteris paribus*, greater weight ought to be assigned to beliefs that are held by many experts rather than a few. See Lehrer and Wagner [1981]. In contrast, Goldman [2001] argues that, because experts judgments might be highly correlated due to common information, agreement cannot always provide greater evidence of a hypothesis. Thus, Goldman claims that there are several heuristics that one might use to evaluate expert testimony.

[2] See Burge [1993], Coady [1973], and Foley [2005] for several different defenses of the non-reductionist position. See Adler [1994], Fricker [1994], Fricker and Cooper [1987] for defenses of reductionism. Additional references are available in Lackey [2011].

vary. These are the aims of this paper.

In Section 1, I develop a formal model of communal learning; the model is most appropriate to understanding dissemination of propositional knowledge in scientific communities. I then use the model to describe six candidate TNs that approximate informal norms such as "believe *p* if it appears to be the majority opinion", and "believe *p* if is endorsed by an expert" and "believe *p* if and only if it appears to be endorsed by a majority of experts." The six TNs resemble rules that are endorsed by reductionists and non-reductionists, and by social epistemologists who advocate "going by the numbers" with respect to expert testimony. To be clear, the six TNs are extremely simple and naive. However, characterizing the reliability of said norms provides a starting point for characterizing the reliability of the more sophisticated TNs that are, at least implicitly, under discussion in social epistemology.

In Section 2, I use the model to characterize those features of scientific communities that influence reliability. I evaluate reliability in three ways: (1) does following a TN lead one to develop true beliefs? (2) if so, how *quickly* does it to the truth?, (3) if error is unavoidable, how often does a TN lead one to believe falsehoods? The final section discusses limitations of my model and directions for future research.

My central findings are as follows. I argue that, in assessing the reliability of TNs, epistemologists ought to play closer attention to miscommunication and the "communicative structure" of scientific communities. Why? In the absence of miscommunication, most TNs are equally reliable in the first two senses above: they lead scientists to develop true beliefs, and they do so at roughly the same speed. Only when miscommunication is present, therefore, can one compare norm reliability. In this case, reliability depends crucially upon the structure of scientific communities. In particular, insular communities, in which scientists communicate most frequently with experts in their own fields, are more reliable in the second sense above but less reliable in the third: they make discoveries more quickly but the accuracy with which such results are disseminated to non-experts is compromised.

My findings are important for epistemology because, in traditional discussions of testimony, the effects of miscommunication and of communicative structure are often ignored, and different senses of reliability are often not distinguished. For philosophers of science, my findings concerning communicative structure illustrate the benefits and costs of current scientific practice, where increased specialization has led to increasingly insular communities. Moreover, the same results also reveal the potential value (or lack thereof) of recent government and university initiatives that attempt to eradicate insularity by sponsoring interdisciplinary research.

1 A Model of Communal Scientific Inquiry

Assume there is some finite set of **questions** that scientists wish to answer. Each question has a set of mutually incompatible **answers**, and scientists aim to find the unique correct answer to each question.[3]

For example, imagine medical researchers are investigating the efficacy of several pills. In this example, each question is of the form "Is pill j effective?", and there are two answers: "yes" and "no." Formally, for each pill j, there is an unknown real number e_j that represents the average effectiveness of the pill, where a pill's "effectiveness" is a function of both its side effects and its efficacy in curing the intended ailment. If e_j is non-negative, then the pill is salutary (on average). Otherwise, the pill is harmful. Moreover, the magnitude of e indicates how harmful or salutary the treatment is. So the formal question is, "Is e_j non-negative or not?"

Assume there are discrete stages of time t_1, t_2, and so on. At each stage, scientists collect **data**. Importantly, data can be misleading in the short-run. In the example, suppose that, at each stage, every researcher treats some fixed, finite number of patients with one of the pills and records the results. Those results are the researchers' data. How can such data be misleading? Imagine that the effect of a pill is probabilistic, and so even if the pill is salutary on average, some patients may react poorly. Thus, it is possible, though unlikely, that a researcher observes forty consecutive patients who react poorly to a pill, even if it is quite beneficial on average.[4]

I assume each scientist has a **specialty**, i.e., her data pertains to exactly one question. Therefore, she must learn the answers to different questions by asking others. This assumption represents the fact that real researchers' abilities are limited due to specialized training, time, and/or financial constraints. In my example, each researcher studies exactly one pill, and she treats patients with that pill only.

To model communication, I represent researchers by nodes in a colored, undirected graph like the one below. The colors indicate researchers' specialties, i.e., two researchers share a specialty if and only if they are represented by nodes of the same color. In my example, I call the pills the "red pill",

[3] Bolded terms are defined in the appendix, which also contains proofs of the theorems.

[4] In computer simulations, I assume there are finitely many pills $1, 2, \ldots, n$. At each stage, a scientist treats a single patient with a pill $i \leq n$ and observes an outcome. Outcomes are normally distributed with unknown mean e_i (i.e. the effectiveness of the pill) and known variance σ_i^2. The normality assumption is immaterial to the results below: similar results are obtained when the agents draw from other types of distributions. Moreover, the theorems do not depend upon assumptions concerning the probabilistic process by which data is generated; in particular, data need not be i.i.d.

"blue pill" and so on. I call a scientist "red" if she studies the red pill.

In my model, two scientists can share information if and only if they are connected by an edge in the graph representing their community. Say two scientists are **neighbors** if they are connected by an edge; a scientist's **neighborhood**, then, can be defined as the set of all her neighbors.

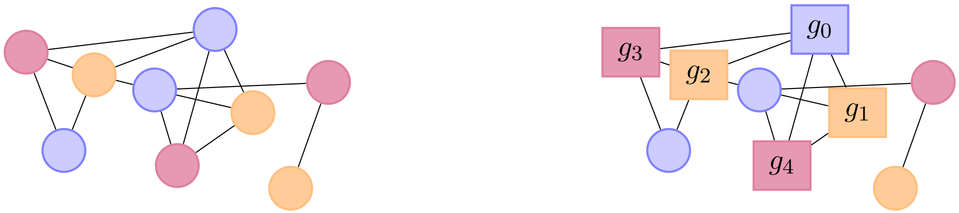


Figure 1: A research network and the neighborhood of g_0 (indicated by squares) in that same network

Not all graphs, however, properly represent scientific communities. For instance, suppose a graph can be divided into (at least) two sections such that information cannot pass (even indirectly through others) from one section to the other. In this case, one should not say that the scientists form a single "community," as different parts of the so-called "community" never interact whatsoever. For this reason, I focus exclusively on **connected** networks, which cannot be divided into two separate parts.

Importantly, I assume the type of information that neighbors can share depends upon their respective specialties. In particular, neighbors with the *same* specialty can share *data*, but those with *differing* specialties can share only their beliefs about the *answers* to questions. In my example, two "red" scientists tell each other how well their patients have responded to the red pill. In contrast, a red and a blue scientist can only ask each other "Do you think the red (or blue, or green, etc.) pill is effective?" and trade answers. Scientists with different specialties cannot even share their quantitative assessments of *how* effective a pill is.

Why assume that researchers can share information in this limited way? In the real world, scientists must rely on the work and findings of others. However, if scientists could always share and evaluate each other's data, then there would be no such reason to rely on others. The assumption that not all *data* is shared, therefore, is intended to capture the fact that certain "high level" judgments (e.g. is the pill effective?) can be communicated easily even if the data and the methodology for evaluating said data cannot.

Why assume that researchers with the same specialty can share data, whereas researchers with different specialties cannot? First, it is generally easier for a scientist to understand the findings, methods, etc. of research conducted in her own field than to understand the work of researchers in remote disciplines. For example, theoretical physicists can (sometimes) competently evaluate articles in theoretical physics, but can rarely understand more than the abstract of a paper in molecular biology. Second, researchers often read only survey articles about work outside their specialties, whereas they often read the articles on which summaries are based within their own fields. Thus, even if a researcher could in principle understand work outside her specialty, she might choose not to do so because of the investment it would require to learn more than what is available in survey articles.

Thus far, I have explained two ways in which a scientist learns in my model, namely, (1) she collects data, and (2) she learn the answers to questions outside her specialty from others. I now explain how my idealized scientists *use* such information to arrive at their beliefs.

Within her specialty, a scientist uses a **method** for inferring answers from data. Importantly, her beliefs within her specialty are determined exclusively by her method and data; they are not influenced by her neighbors' beliefs. Formally, a method is a function from data sequences to answers. I assume that each scientist's method is **convergent**, i.e., whatever the truth, with probability one, there is some stage at which the method eventually conjectures the truth and does so from that stage onward.

In my example, each scientist employs a method such that, whether her pill is effective or not, there is some stage at which she conjectures the pill is effective if and only if it is so. Specifically, each researcher employs a statistical test to determine whether her particular pill is effective or not.[5] This part of my model mirrors scientific practice closely, as statistical tests are the trade of most medical researchers and social scientists. The way in which a scientist learns answers to questions *outside her specialty* is more complex and is explained in the next section.

Three caveats are necessary. First, I assume researchers' methods find the truth *in their specialties*; no assumptions thus far guarantee finding the truth outside one's specialty. Second, researchers do not know when the truth has been discovered. In my example, a scientist may correctly believe her pill is effective, but it is possible that her data are misleading. So future

[5] In simulations, scientists employ likelihood ratio tests (LRTs) to test the null hypothesis $e_i \geq 0$ vs. the alternative $e_i < 0$, where e_i is the effectiveness of pill i. To make such tests convergent (specifically, s.a.s convergent as defined in the appendix), significance levels are adjusted downward with sample size. Again, the use of LRTs, or frequentist rather than Bayesian methods, is inconsequential for the analytic results below. What matters is that scientists methods are convergent.

data might undermine her current belief.

Third, methods are not guaranteed to be *quick*. For example, suppose a scientist conjectures her pill is effective if at least half of her patients have positive outcomes. Now suppose that the pill is effective, but that the scientist's first 100 patients react poorly. Such outcomes are unlikely but possible. Then the scientist's method will lead her astray until she sees many positive outcomes, which will take at least 100 more observations. In general, I assume that a scientist's method *eventually* returns the true answer, but, there may be no positive number n - no matter how large - such that one is guaranteed true belief after n stages.

1.1 Testimonial Norms

Recall, a **testimonial norm** (TN) is a rule for accepting or rejecting the claims of others. In my model, an agent's TN dictates which answers she believes to questions outside her specialty. In this section, I describe six simple TNs. Although the six norms are motivated by debates in social epistemology, no philosopher, to my knowledge, has endorsed the norms below. However, studying these six can shed light on more complex TNs, and many of the results below show that even naive norms are reliable.

Assume that, on each stage, researchers must decide which answer to believe to questions outside their specialties. I call an agent a **Reidian** if she adopts the opinion of a randomly chosen neighbor. I call her a **majoritarian Reidian** if she adopts the opinion of the majority of her neighbors.

Notice both types of Reidians ignore their neighbors' expertises. For example, a Reidian may consult red expert about the green pill even if she has a green neighbor. In real-world settings, however, individuals often attribute more weight to the opinions of experts, and such reliance on intellectual credentials is often thought to be rational.

To model reliance on experts, I call an agent a **e-truster** if she adopts the opinion of a random expert neighbor if she has one, and otherwise, trusts a random neighbor. For instance, when a blue e-truster is deciding whether the red pill is effective, she asks one of her red neighbors; if she has no red neighbors, then she asks a random neighbor. A **majoritarian e-truster** adopts the opinion of the majority of her expert neighbors; if she has no expert neighbors, she adopts the opinion of the majority of her neighbors. Thus, e-trusters (majoritarian or not) distinguish experts from non-experts. However, they treat all experts as equally reliable.

Alvin Goldman and others have argued that the norm of e-trusting is unreliable: they claim that agents ought to assess the reliability of experts to determine whom to trust. According to [Goldman, 2001], there are various heuristics for evaluating reliability. For example, one might attribute greater weight to the opinions of an expert who advances cogent arguments. Unfortunately, not all of the Goldman's heuristics can be accurately captured in my model. Here, I restrict myself to modeling one heuristic, which I call **informational proximity**.

The informal concept of informational proximity is best illustrated by examples. I am not a physicist, nor do I interact with physicists. However, some of my colleagues, who study philosophy of physics, do interact with physicists. Those colleagues, therefore, are more informationally proximate than I am to current work in particle physics, say. Hence, if a student is deciding whether to accept my testimony or that of a philosopher of physics when it concerns current work in particle physics, she might consider the latter to be more reliable the former because of the informational proximity of philosophers of physics to the facts.

To model informational proximity, define the **distance** between two researchers to be the shortest path in the undirected graph representing their scientific community. Notice both Reidians and e-trusters ignore informational proximity. For example, suppose an e-truster has two neighbors, neither of which is an expert in the question q. One of her neighbors, however, is connected to a q-expert, whereas the other is three-degrees-removed from a q-expert. Real-world scientists might favor the former's opinion, as the former is closer to the source of reliable information. In contrast, e-trusters in my model ignore informational proximity when they adopt the opinion of a random neighbor.

Call a researcher a **proximitist** if, on any given stage of inquiry, she adopts the opinion of the neighbor who is closest to a q-expert when deciding which answer to q to believe; if there are multiple such neighbors, then she chooses one at random. Call an agent a **majoritarian proximitist** if she polls her most proximate neighbors.

Examples of TNs can be multiplied indefinitely. However, the six considered here are important because they differ on several dimensions that have been the focus of debates in social epistemology. By contrasting Reidianism, e-trusting, and proximitism with their majoritarian counterparts, one can investigate the consequences of "going by the numbers" versus those of reliance on one individual. And although no non-reductionist may endorse Reidianism, one can investigate the value of seeking positive reasons to trust a speaker (by employing heuristics like informational proximity) by comparing Reidianism, e-trusting, and proximitism. Perhaps surprisingly, it turns out that Reidians (though not majoritarian Reidians) reliably acquire true beliefs in the absence of miscommunication. To see why, I compare the reliability of the six norms, among others, in the next section.

2 Reliability

2.1 Convergence

One way to evaluate the performance of TNs is to investigate which are truth-tracking. Formally, say a TN is **convergent** if, whatever the truth about the world, when a network adopts said norm, every researcher will hold only true beliefs given some (potentially large) finite amount of data.[6] Unfortunately, convergence is insufficient to distinguish among four of the norms by the following theorem:

Theorem 1 *In connected research networks, Reidianism, e-trusting, proximitism, and majoritarian proximitism are convergent. In contrast, majoritarian Reidianism and/or majoritarian e-trusting are not.*

In fact, there is nothing special about the four convergent norms. The above theorem can be generalized to show that every norm satisfying basic requirements of rationality and realism is convergent.[7] The philosophical upshot is that, if reliability is understood as the eventual acquisition of true belief, then there is no difference among a wide class of norms. In particular, the decision to adopt a "reductionist" norm, which might require one to find positive reasons to trust a speaker, versus a "non-reductionist" norm, which might permit one to trust others in the absence of defeating conditions, is unimportant as long as the norms satisfy basic normative constraints.

[6] The notion of convergence employed here is what I call "strong almost-sure convergence" which is stronger than almost-sure convergence (and hence, convergence in probability) in general, but logically equivalent to almost-sure convergence when the partition of the parameter space is finite. See the appendix for details.

[7] See the definitions of finite memory, stability, and sensitivity in the appendix. Any realistic TN has finite memory, and though space prevents me from arguing so here, I believe stability and sensitivity are minimal normative requirements for any TN. Majoritarian Reidians do not converge because they can get caught in "echo chambers": if a tightly-connected group of agents all have identical, false beliefs about the efficacy of some pill outside their area of expertise, then when each agent in the group polls her neighbors, she will find her opinion to be in the majority and stick to it. So each agent in the group holds some belief precisely because others in the group do. Because agents in the group polls all neighbors, and not just the experts, it follows that the group's beliefs cannot be penetrated by external information from experts. Majoritarian e-trusters fail to converge because they behave like majoritarian Reidians in the absence of expert neighbors.

The above theorem, however, neglects the *speed* with which agents learn. One might wonder, "if reliability is understood in terms of *quick* convergence, are there any differences among the four convergent norms?" Surprisingly, the answer is "no." To see why, define **convergence time** to be the number of stages elapsed before every researcher holds true beliefs and will continue to hold such beliefs indefinitely. Thus, the second way of evaluating various norms is to consider the question, "which norms minimize average convergence time?"

To answer this question, I simulated the running example of my model.[8] First, I randomly generated approximately 4500 graphs consisting of between 50 and 100 agents. Disconnected graphs were removed from the data because no norms are convergent in disconnected networks. To model the fact that communication is limited, I generated graphs in which researchers were neighbors with at most 10% of the other agents.

Equal numbers of agents were assigned one of five specialties, and the network was assigned one of the four convergent norms. A simulation was stopped when all agents' beliefs were true for ten consecutive stages, and the tenth to final stage was assumed to be the convergence time of the network.

What effect(s) do norms have on convergence time?[9] The answer: essentially none. Except in the "easiest" problems, there is no statistically significant relationship between norms and convergence time: populations of Reidians, proximitists, and so on, all converge at the same rate on average. In "easy" problems, Reidians converge at a rate slower than the remaining norms on average, but there is no significant difference among the remaining three convergent norms.

Although these results may seem surprising, I claim they are intuitive and illustrate a robust pattern in the history of science. Consider any difficult scientific undertaking - for example, understanding the principles of flight. Before the Wright brothers, humans had attempted to engineer planes for millennia. So the discovery of principles of flight took at least a few thousand years. In contrast, once the first planes had been constructed, the engineering knowledge spread worldwide in only a few years. The time to *disseminate* such knowledge, therefore, was minuscule in comparison to the time it took to *discover* it.

In general, when scientists are faced with a difficult question, discovery is slow. However, the time required to communicate their eventual findings

[8] The code for the simulations can be found on the author's website.

[9] Convergence times were compared using a one-way random effects analysis of variance. The raw data and relevant sample statistics are available on the author's website.

to non-experts may be the same as if the question had been easy. This explains why TNs have no significant effect on convergence time in my model. Whereas methods are responsible for discovery, TNs affect only the speed of dissemination. As the questions become more difficult, the time required for dissemination is dwarfed by discovery time. Hence, TNs have only a negligible effect on convergence time when questions are difficult.

In sum, infinitely many TNs are convergent, and when scientific questions are difficult, there is no significant difference among TN convergence speed. Since science is a difficult enterprise, one might conclude that choice of TN is irrelevant. This conclusion is hasty. Although there are several idealizations in my model, two deserve greater scrutiny.

First, I have assumed that that agents never misspeak or misinterpret others. Does a TN's reliability change when miscommunication is possible? Second, I have evaluated TNs in *all* possible networks, including those that do not represent real world scientific communities. Do the relative performances of TNs change in more realistic networks? These two questions are the subject of the next two sections.

2.2 Miscommunication

Miscommunication is an unavoidable feature of human interaction. Speakers make errors that result in ambiguity and/or unintended meanings, and listeners may misinterpret what speakers say. Misunderstandings seem fairly common in academic communities when researchers in one field try to share their findings with others with radically different knowledge and training. Does such miscommunication affect the reliability of various TNs?

To answer this question, I investigate the effects of miscommunication within the running example. I assume there is fixed (i.e., for all time) probability $\epsilon < \frac{1}{2}$ of miscommunication. That is, if g believes the red pill is effective and a neighbor n asks g her opinion, then n will believe that g reported the red pill to be *in*effective with probability ϵ.[10] As in the previous section, one can ask, "which TNs are convergent when miscommunication is present?" The answer: essentially none.

Why does miscommunication prevent convergence? Suppose that - as is the case in the real world - researchers remember only part of the past, and their beliefs depend only upon on what they remember. Formally, say a TN has **finite memory** if there is a number n such that the TN is a function of neighbors' beliefs on only the last n stages. The six TNs above all have finite

[10] None of the results below rely on the fact that ϵ is the same for all agents. This assumptions is made for simplicity of calculations and proofs only.

memory (of length one!). Now, if there is some fixed probability (however small) of miscommunication, then there is always some chance that, even when all of her neighbors hold true beliefs, an agent will misunderstand their claims for as many stages as she can remember. So, if agents have finite memories, then even the most ingenious TNs will fail to converge:

Theorem 2 *Suppose there is some fixed, non-zero probability of miscommunication. Then no* TN *with finite memory is convergent. In particular, none of six* TN*s considered above is convergent.*

The theorem only shows that agents might *occasionally* believe false statements. Perhaps it is wrong, then, to demand that TNs converge. Rather, one should be interested in TNs that *minimize error*.

How should one calculate error? Imagine taking a snapshot of all agents' beliefs on a given stage. Given the snapshot, one can calculate the proportion f_n said beliefs that are false. So given a network, one can (in theory) calculate the *expected* number of false beliefs $e_n = E[f_n]$ on stage n. Call e_n the **error rate on stage** n. For many TNs, e_n fluctuates wildly from one stage to the next. Luckily, the six TNs above are not of this sort:

Theorem 3 *Suppose there is some fixed probability of miscommunication. Then each of the six* TN*s above converges to some fixed error rate. That is, e_n approaches a fixed value e as n approaches infinity. If the probability of miscommunication is non-zero, then the error rate is positive.*

Call the fixed value e the **error rate** of the network. Hence, one can compare the reliability of TNs by comparing their error rates. Simulation results show that, for all problem difficulties and all networks, the error rate of Reidians is on average greater than that of e-trusters. E-trusters err more often than do proximitists, who in turn, err more than majoritarian proximitists.[11] However, although Reidians err more often than do e-trusters and proximitists, their respective error rates depend crucially upon "network structure." This is the subject of the next section.

2.3 Network Structure

Thus far, I have argued that, if communication is perfect and one's goal is the quick acquisition of true beliefs, then choice of TN is irrelevant. In contrast,

[11] Statistical tests supporting these claims are summarized in the appendix to the longer version of this paper.

if miscommunication is present, then different TNs have differing error rates. In analyzing the simulation results, however, I have made no attempt to distinguish realistic network structures from mathematically possible, but highly unrealistic, ones. Are there any features of real scientific communities that might affect the performance of TNs?

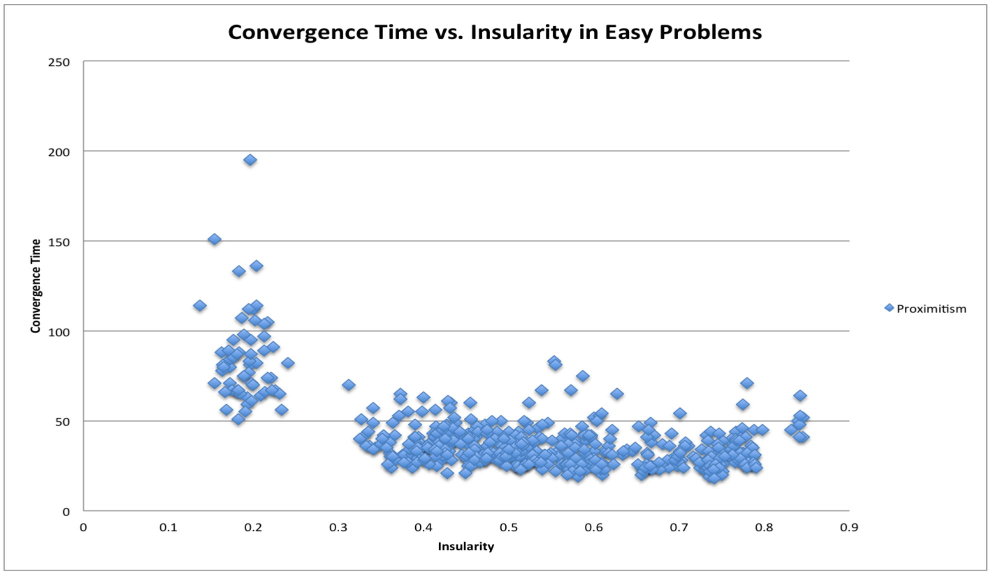
One way in which academic communities are unique is that they are divided into *research units*. Roughly, a research unit is a collection of individuals who (i) have similar research programs and (ii) communicate with one another frequently. Sometimes, a research unit is a lab or an academic department at a university. Other times, research units are comprised of academics who live in different parts of the world, but still read each others' papers, collaborate, and so on.

In my model, research units are represented by groups of agents who share a specialty and are highly connected. Formally, for any agent g, define **g's insularity** to be the proportion of her neighbors who share her specialty. Define a network's **insularity** to be the average insularity of its agents.[12]

Intuitively, insularity seems both desirable and dangerous. On one hand, researchers with the same expertise ought to communicate as frequently as possible. On the other hand, when academic communities become too insular, there is a chance that one research unit completely isolates itself, thereby failing to share its own findings or draw upon the work of others. So too much insularity is harmful.

These intuitions are captured by my model. For the moment, I ignore miscommunication once again. Below is a graph indicating that proximitists converge more quickly (on average) in more insular networks; similar results are obtained for the other three convergent norms.

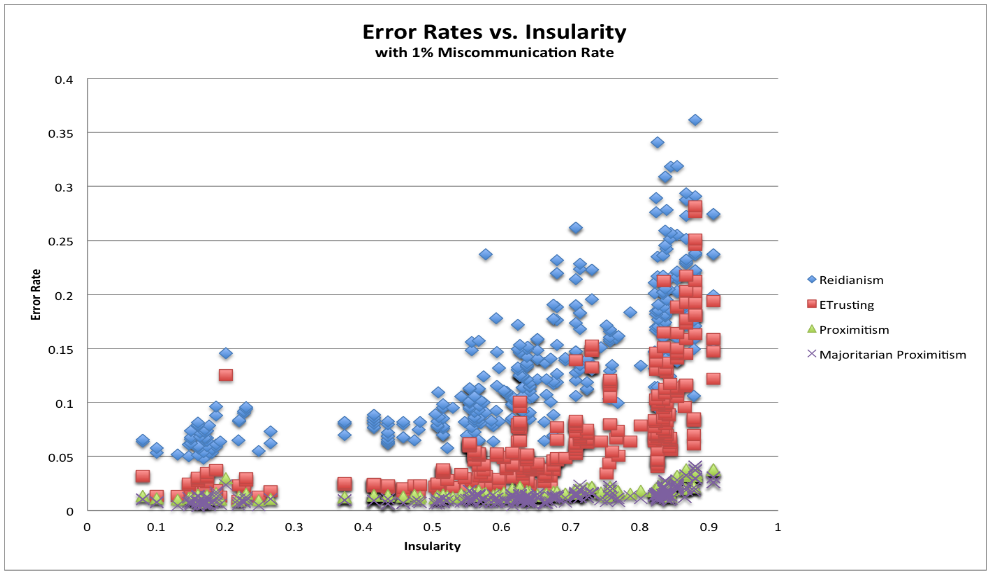
[12] What I call insularity is often called **homophily** by social scientists. There is relatively little work about homophily affects learning. An important exception is Golub and Jackson [2012], which develops a radically different model from the one presented here.



The reason that insularity decreases convergence time is fairly simple. Recall that scientists with the same expertise can share data in my model. More data allows a researcher to ascertain the true answer in her specialty more quickly, and therefore, decreases convergence time.

When miscommunication is present, a similar result holds when one considers the time it takes a network *to converge to its error rate*. Because agents employ convergent methods, the error rate of a network is determined entirely by the number of false beliefs agents hold with respect to questions outside their respective specialties. During the discovery stage, the frequency of false beliefs will typically be higher than the (asymptotic) error rate. Why? By definition, during the discovery stage, agents may have false beliefs in their own specialties. It follows that quickening discovery shortens the time until a network converges to its error rate. Since insular networks have shorter discovery stages, they will also typically converge to their error rates more quickly.

Above, I claimed that insularity is both desirable and dangerous. The graph above shows it to be desirable. What is its danger? The answer: higher error rates. Below is the graph that shows that, for each of the four convergent TNs, more insular networks typically have higher error rates. However, the rate at which error rates increase differs among the four GTNs. In the presence of miscommunication, both radical and e-trusters quickly become unreliable as insularity increases, whereas both proximitists and majoritarian proximitists have much slower growing error rates.



The reason that insularity increases the error rates of the four TNs s fairly easy to explain. Error rates are a consequence of a "telephone-game effect". When an agent learns a fact first-hand from an expert, there is only a small chance of miscommunication. When an agent learns a fact second-hand, the chance that miscommunication has occurred is higher: not only is there a chance of miscommunication between an agent and her informant, but also there is a chance of miscommunication between the informant and the expert from which the informant learned the fact. So as agents become more distant from experts, the chance that miscommunication has occurred increases. When a network is insular, informational paths between two experts in different fields are generally longer, and hence, error rates are typically higher. This suggests that to minimize error rate and ensure high speeds of convergence, ideal networks ought to balance insularity and average path-length between agents; so-called "small worlds" networks often have this property precisely.

In sum, insular networks converge to (mostly) true belief more quickly, but there is trade-off between speed of learning and error.[13] It is an open question whether there are any TNs for which error rates *decrease* as insularity increases.

[13] Zollman [2011] finds a very similar trade-off between speed and reliability in a different model of scientific inquiry. This is evidence that the phenomenon (i.e., the trade-off) is robust under varying modeling assumptions.

3 Conclusions and Future Research

I have argued that, in assessing the reliability of TNs, epistemologists ought to pay closer attention to miscommunication and the communicative structure of scientific communities. My argument went as follows. In the absence of miscommunication, most TNs are equally reliable in two senses: they lead to true belief eventually and do so at roughly the same speed (by Theorem 1 and simulation results). In the presence of miscommunication, no TNs are reliable in either of these senses (by Theorem 2). Luckily, reliability can be compared in the third way: different TNs may have differing error rates (by Theorem 3). Simulation results indicate that reliance on experts decreases error, but error rates depend crucially upon the structure of communities. In particular, insular networks, in which researchers communicate primarily with similar specialists, make discoveries more quickly but at the cost of less accurate dissemination of said discoveries to non-experts.

The above argument assumes that my idealized model can be used to draw conclusions about real scientific communities. In the remainder of the paper, I will discuss three idealizations; doing so clarifies the range of applicability of my model and suggests questions for future research.

First, in my model, areas of expertise are unrelated: one scientist's findings are useless to researchers with different specialties. In real scientific communities, physicists' models can be applied to economic phenomena; economists' techniques are useful in biology, and so on. A more realistic model, therefore, should represent the complex collaborative relationships among different academic disciplines.

These considerations suggest ways in which my model might be extended, but one should be careful not to infer that they make my model completely inapplicable to science. Although I have suggested that specialties might represent academic disciplines, one need not interpret my model this way. Two scientists may study a similar *question* in my sense, yet they may work in different *disciplines*. Thus, because researchers with the same specialty do collaborate in my model, there's nothing, in principle, that prevents my model from representing interdisciplinary collaboration.

Moreover, while science is often collaborative, much research is also carried out in parallel. For example, molecular biologists and sociologists work in parallel, as techniques for copying DNA need not inform research in military sociology or vice versa. My model, at the very least, captures this type of parallel research. Nonetheless, future research ought to investigate the reliability of TNs when researchers cannot answer their questions in parallel and in which specialties (in my sense) correspond to academic disciplines.

Second, in my model, network structure is *static*. However, real scientific communities change: older scientists die and others enter the profession; new collaborations are born while others fade, and so on. Preliminary simulation results suggest that that many of the above results hold even if network structure changes over time (e.g., many TNs converge in the absence of miscommunication; they converge to an error rate in the presence of miscommunication, and so on). Nonetheless, dynamic scientific communities raise a number of new questions. How should the underlying graphical structure representing scientific communities evolve to mirror the real-world dynamics of scientific communities? Is there a way to extend the concepts of "insularity" and "informational proximity" to dynamic networks?

Finally, in my model, agents exchange answers to questions without providing reasons for their opinions. A more realistic model might represent the exchange of *arguments*. Of course, this idealization is both a virtue and vice of my model. Researchers in the real-world often lack the ability to critically evaluate the intricate arguments of scientists in other fields, and my model aims to capture this fact.

Nonetheless, there are also circumstances in which researchers with differing specialties can competently evaluate each others' arguments. How one should define and model TNs for argumentative agents remains an open question, but it is crucial to investigating the reliability of the more realistic TNs, especially those that use various heuristics (like those suggested by Goldman) for evaluating expert reliability.

4 Acknowledgments

This project was inspired by Kevin Zollman's work on the communicative structure of scientific communities and, in particular, by a presentation he gave at the 2012 Pitt-CMU graduate conference on testimony. I am thankful to Kevin thanks for both his comments on earlier drafts of this paper and for intellectual inspiration. I am also indebted to Gerhard Schurz for suggesting that I investigate the effects of miscommunication within my model.

Thanks to Jan Sprenger, David Danks, Clark Glymour, Kevin Kelly, and Michael Strevens, who all provided detailed comments, constructive criticism, and helpful suggestions for improving earlier drafts of this paper. Finally, the paper also benefited from comments by an anonymous referee and participants at the "Collective Dimensions of Science" conference in Nancy, France.

This work was partially supported by the National Science Foundation under Grant No. SES 1026586. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Jonathan E. Adler. Testimony, trust, knowing. *The Journal of Philosophy*, 91(5):264—275, 1994.

Tyler Burge. Content preservation. *The Philosophical Review*, 102(4):457—488, 1993.

C.A.J. Coady. Testimony and observation. *American Philosophical Quarterly*, 10(2):149—155, 1973.

Richard Foley. Universal intellectual trust. *Episteme: A Journal of Social Epistemology*, 2(1):5 — 11, 2005.

Elizabeth M. Fricker. Against gullibility. *Synthese Library*, pages 125—125, 1994.

Elizabeth M. Fricker and D.E. Cooper. The epistemology of testimony. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 61:57—106, 1987.

Alvin I. Goldman. Experts: Which ones should you trust? *Philosophy and Phenomenological Research*, 63(1):85—110, 2001.

Benjamin Golub and Matthew O. Jackson. How homophily affects the speed of learning and best-response dynamics. *Forthcoming in Annals of Economics and Statistics*, 2012.

Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability.* American Mathematical Society, 1997.

Jennifer Lackey. Testimony: Acquiring knowledge from others. *Social Epistemology: Essential Readings, ed. Alvin I. Goldman & Dennis Whitcomb. Oxford: Oxford University Press*, pages 314—337, 2011.

Keith Lehrer and Carl Wagner. *Rational Consensus in Science and Society: A Philosophical and Mathematical study*, volume 24. D. Reidel, 1981.

Kevin J. Zollman. The communication structure of epistemic communities. In A Goldman, editor, *Social Epistemology: Essential Readings*, pages 338—350. 2011.

5 Appendix

5.1 General Strategy

The theorems in the text are all consequences of well-known results about Markov processes. Here, I explain the general strategy of the proofs, as the notation below is rather cumbersome. To do so, I first informally sketch a proof of Theorem 1, which asserts that four of the testimonial norms in the body of the paper converge in the absence of miscommunication.

Consider a network of Reidians. Imagine one of the Reidians - let's call her Jane - is deciding which of her neighbors to trust at time t concerning a question outside her area of expertise. If Jane has an expert neighbor, Jill, then there's some chance that Jane will adopt the Jill's opinion. At time t+1, there's some chance that Jane's neighbors will adopt Jane's beliefs. Then at time $t + 2$, neighbors of neighbors of Jane may adopt Jane's belief, and so on. In this way, there's some chance that Jill's expert opinion propagates through the entire network (if the network is connected), and this is true at any point in time. So there's some chance that every agent outside of Jill's area of expertise will eventually hold Jill's belief.

Since experts eventually hold true beliefs in their area of expertise, this entails that all agents will eventually hold Jill's *true* belief. Once everyone has the same, true belief, the Reidian norm ensures that everyone continues to believe it. Since this is true of every area of expertise, the network must converge. Similar arguments apply to e-trusters, proximitists, and majoritarian proximitists.

This argument is an instance of a more general argument concerning absorbing Markov processes. In general, **Markov process** is one which a system's behavior depends only upon its most recent past. For example, consider a gambler's total earnings. Suppose a gambler repeatedly makes $20 bets on a roulette wheel. Then the gambler's total earnings at time $t + 1$ depend only upon her earnings at time t (and whether or not she wins her current bet); whether the gambler was nearly bankrupt or enormously wealthy at some stage prior to t is irrelevant to her total earnings at $t + 1$.

Some Markov processes have **absorbing states**, which means that the system reaches a state that it cannot leave. For example, suppose our gambler is very unlucky and loses all of her money at time t. Then she cannot make any more bets, and so her total earnings equal zero from time t onward. For this reason, the state of having no money is called absorbing. Similarly for the state in which the gambler wins all of the casino's money.

Under very weak conditions, a Markov process with absorbing states will eventually transition into one of its absorbing states and stay there forever. The crucial assumption is that there is some number n such that, no matter what state the Markov process is, there's some finite probability that it will transition to an absorbing state in n steps. This guarantees that the probability that the process transitions to an absorbing state at some point is one. To prove that the TNs in the paper are convergent, therefore, it suffices to show three assumptions are met: (1) agents' beliefs behave like a Markov process, (2) the state in which they all have true beliefs is uniquely absorbing, and (3) there is some number n such that, for any state of the process, there is positive probability that agents will find themselves in the unique absorbing state (of all true beliefs) n stages in the future.

For the first assumption, note that all the testimonial norms in the body of the text have a common property: an agent's beliefs at time $t+1$ (outside her area of expertise) depend only upon what her neighbors believed at time t. For example, a Reidian fixes her belief at time $t+1$ by adopting a neighbor's belief at time t. So her neighbors' beliefs prior to t, no matter how steadfast or erratic, are completely irrelevant to what she believes now. Consider the vector of all agents' beliefs about all pills under investigation. If all the agents employ TNs like the ones in the text, therefore, their beliefs will behave like a Markov process, just like the gambler's earnings.

The argument above is right in outline, but a tiny bit too fast. Recall, agents' opinions at a given time t depend not only upon others' opinions, but also upon data. According to my assumptions, agents can use methods to make inferences from any amount of data whatsoever, not just the set of most recent observations. So technically, agents' beliefs do not form a Markov process. This is, however, easily fixed.

Recall, I assume that for every researcher in the network, there is some stage of inquiry at which she holds true beliefs *in her area of expertise* from that stage onward. Let E_n be the event that all researchers have converged to the truth in their respective domains by stage n. The crucial observation is that, *conditional on* E_n, the evolution of agents' beliefs do form a Markov process. Why? Researchers beliefs in their area of expertise are fixed from n onward by assumption, and by definition, the six TNs make one's beliefs outside of one's area of expertise dependent only upon the last stage. Since experts beliefs converge to the truth with probability one, it follows that some E_n will occur with probability one, and so the evolution of agents' beliefs will behave like a Markov process after enough time. This is what the first two lemmas concerning what I call **PC-Markov processes** say below. So much for the first assumption.

What about the second assumption? Notice that agents employing the TNs in the text only change their beliefs if their neighbors disagree. So it looks like any state in which all agents have identical beliefs is an absorbing one. However, "all true beliefs" is the unique absorbing state as I assumed that experts eventually have true beliefs in their area of expertise. In general, any TN satisfying what I call **stability** will be absorbed in such a way; stability says that if an agents' neighbors unanimously believe φ for as long as the agent can remember, then the agent also believes φ.

Finally, one needs to show that the network transitions to the absorbing state of "all true beliefs" with probability one. This is what the Jane and Jill example showed. For Reidians and e-trusters, there is some positive probability that experts' opinions propagate through the network in at most L many steps, where L is the length of the longest path in the network. In general, an assumption that I call **sensitivity** guarantees that true expert opinions propagate through the network. Roughly, an agent's TN is sensitive if, for every area of expertise, there is some positive probability that the agent adopts the belief of her neighbors that are closer to an expert in the area, regardless of what others in her neighborhood believe. Proximitism is sensitive because agents always trust their more proximate neighbors; Reidianism and e-trusting are sensitive because they trust all their neighbors with some positive probability. Majoritarian Reidianism and majoritarian e-trusting are not sensitive because neighbors who are closer to an expert can be outvoted.

In the presence of miscommunication, however, the "all true beliefs" is no longer absorbing: an agent may misinterpret her neighbors and develop false beliefs, even if her neighbors' beliefs are all true. Instead, agents beliefs evolve according to a **regular** Markov process (again, conditional on having converged in their respective areas of expertise). Regular Markov processes are, in a sense, the opposite of absorbing ones: there is some number of steps n such that the process can transition from any state to any other state in exactly n steps. No state is absorbing. It can be shown that, for each state s of a regular Markov processes, the probability that the process is in state s approaches a fixed value as the number of states gets large.

Why do agents' beliefs evolve according to a regular Markov process in the presence of miscommunication? The six TNs are open-minded in the sense that for any particular pill and any judgment about the pill (i.e.

effective or not), there is some vector n such that if an agent thinks her neighbors beliefs are represented by n, then she will adopt the judgment in question about the pill. Let b be any vector describing all agents' beliefs outside their respective areas of expertise. Because there is some fixed, positive chance of miscommunication on any stage of inquiry, there is some positive probability that agents will hear exactly what they need to hear from their neighbors to form beliefs b, regardless of what everyone in the network actually believes at the moment. So agents beliefs' can always transition to b in exactly one step, which means the process is regular. Note, because b was arbitrary, it also follows that agents always have some positive chance of developing false beliefs, regardless of whether experts have converged to true ones.

This immediately entails that the error rate of the network approaches a fixed value as Theorem 3 says. Let $err_t(b)$ be the number of erroneous beliefs in the network if agents beliefs' are represented by b and if the truth is t. Because agents' beliefs evolve according to a regular Markov process, the above mentioned theorem entails there is some fixed probability $p(b)$ that agents will have beliefs b in the limit. The error rate of the network is just the weighted average $\sum_{b \in B} p(b) \cdot err_t(b)$, where B is the set of all possible belief vectors for the network.

5.2 Notation

Let S^T denote all functions from T to S, and define $S^{<\mathbb{N}}$ to be all finite sequences from S. Let $|S|$ denote the cardinality of S; when S is a sequence, $|S|$ is therefore its length. Given a sequence σ and $n \leq |\sigma|$, let σ_n denote the n^{th} coordinate of σ. If the coordinates of σ are likewise sequences, then let $\sigma_{n,k}$ be the k^{th} coordinate of the n^{th} coordinate of σ. And so on. Let $\sigma \upharpoonright n$ denote the initial segment of σ of length n.

Given sets $S_1, S_2, \ldots, S_n$, let $\times_{j \leq n} S_j$ be the Cartesian product. Given a collection of σ-algebras $\langle S_i, \mathcal{S}_i \rangle_{i \in I}$, let $\otimes_{i \in I} \mathcal{S}_i$ denote the product algebra. In particular, $\otimes_{n \in \mathbb{N}} \mathcal{S}$ is the infinite product space generated by a single σ-algebra. Given a σ-algebra $\mathcal{S}$, let $\mathbb{P}(\mathcal{S})$ denote the set of all probability measures on $\mathcal{S}$. If $p \in \mathbb{P}(\mathcal{S})$, let $p^n \in \mathbb{P}(\otimes_{k \leq n} \mathcal{S})$ denote the product measure on $\otimes_{k \leq n} \mathcal{S}$. When $\mathcal{S}$ is a Borel algebra, these measures extend uniquely to a measure p^∞ on $\otimes_{n \in \mathbb{N}} \mathcal{S}$ (i.e., $p^\infty \in \mathbb{P}(\otimes_{n \in \mathbb{N}})$ is the unique measure such that $p^\infty(F_1 \times F_2 \ldots \times F_n \times S^{\mathbb{N}}) = p(F_1) \cdot p(F_2) \cdots p(F_n)$, where $F_i \in \mathcal{S}$ for all $i \leq n$). Given a metric space M, let $\mathbb{B}(M)$ denote the Borel algebra.

5.3 Preliminaries

The appendix assumes familiarity with Markov processes. For definitions of undefined terms below, see any introductory exposition of Markov processes; I will refer to Chapter 11 in Grinstead and Snell [1997].

Consider a sequence $\langle X_n, E_n \rangle_{n \in \mathbb{N}}$, where the X_n's are random variables and the E_n's are events. Call the sequence a **piecewise conditional Markov process** (or pc-Markov process) if

1. The events $\langle E_n \rangle_{n \in \mathbb{N}}$ are pairwise disjoint,
2. $p(\cup_{n \in \mathbb{N}} E_n) = 1$, and
3. $\langle X_k \rangle_{k \geq n}$ is a time-homogeneous Markov process with respect $p(\cdot | E_n)$, i.e., $p(X_{k+1} | E_n, X_1, X_2, \ldots X_k) = p(X_{k+1} | E_n, X_k)$ for all $k \geq n$.

Call $\langle X_k \rangle_{k \geq n}$ the **pieces** of a pc-Markov process, where X_k is interpreted as a function from a probability space with the measure $p(\cdot | E_n)$. Say a pc-Markov process is **uniform** if all its pieces have the same transition matrix.

Say a pc-Markov process is ergodic/regular/absorbing if each of its pieces is ergodic/regular/absorbing. Say it is **uniformly regular/absorbing** if it is uniform and the pieces are regular/absorbing. The next two theorems show that the asymptotic behavior of uniformly absorbing (or regular) pc Markov processes is identical to that of each of their pieces. They are routine corollaries of 11.3 and 11.7 in [Grinstead and Snell, 1997] respectively.

Theorem 4 *Suppose $\langle X_n, E_n \rangle_{n \in \mathbb{N}}$ is a uniformly absorbing, pc-Markov process with absorbing states S_*. Then $p(\lim_{n \to \infty} X_n \in S_*) = 1$.*

Theorem 5 *Suppose $\langle X_n, E_n \rangle_{n \in \mathbb{N}}$ is a uniformly regular, pc-Markov process with transition matrix $\boldsymbol{P}$. Then for any state $s_i \in S$ there is some probability $r_i \in [0, 1]$ such that $\lim_{n \to \infty} p(X_n = s_i) = r_i$.*

5.4 Definitions

5.4.1 Worlds and Questions

Define a **question** to be a triple $\langle W, \langle \Theta, \rho \rangle \rangle$, where W is a set called **worlds**, Θ is a partition of W, and ρ is a metric on Θ. Elements of Θ are called **answers**. Given a world w, let $\theta_w \in \Theta$ be the partition cell containing w.

Example: Suppose one is interested in determining whether the mean of a normal distribution is at least zero. In this case, W is the set of ordered pairs $\langle \mu, \sigma^2 \rangle \in \mathbb{R} \times \mathbb{R}^+$ representing the mean and variance of the unknown distribution, and $\Theta := \{\theta_{\geq 0}, \theta_{<0}\}$, where $\theta_{\geq 0} = \{\langle \mu, \sigma^2 \rangle \in W : \mu \geq 0\}$ and $\theta_{<0} = \{\langle \mu, \sigma^2 \rangle \in W : \mu < 0\}$. Define ρ to be the discrete metric on Θ. Let Q_N be the question described here. Q_N is the question described in the example in the body of the paper.

5.4.2 Learning Problems and Methods

A **data generating process** for a question $Q = \langle W, \langle \Theta, \rho \rangle \rangle$ is a pair $\langle \langle D, \mathcal{D} \rangle, c \rangle$ where $\langle D, \mathcal{D} \rangle$ is a measurable space, and $c : W \to \mathbb{P}(\otimes_{n \in \mathbb{N}} \mathcal{D})$ is a function, whose values c_w are called the **chances** under w. A **learning problem** L is a pair consisting of a question and a data generating process. Informally, D represents **data**. For all $w \in W$, the probability measure c_w specifies how likely one is to observe particular data sequences.

Example: Let $Q = Q_N$, $D = \mathbb{R}$ and $\mathcal{D} = \mathbb{B}(\mathbb{R})$. For every world $w = \langle \mu, \sigma \rangle \in \mathbb{R} \times \mathbb{R}^+$, let p_w be the unique measure on $\mathcal{D}$ such that the density of p_w is a normal distribution with mean μ and variance σ^2. So data are just sample points pulled from a normal distribution. Let $c_w = (p_w)^\infty$, and let L_N be learning problem described here.

A **method** for a learning problem L is a function $m : D^{<\mathbb{N}} \to \mathbb{P}(\mathbb{B}(\Theta))$. Let $m_d := m(d)$ for all $d \in D^{<\mathbb{N}}$. Informally, a method takes data sequences as input and returns sets of answers with different probabilities.

Example: In the learning problem L_N, one method is to employ a likelihood ratio test to test the null hypothesis $H_0 : \mu \geq 0$ versus the alternative, where the significance of the test is decreased at a rate of the natural log of the sample size. Formally, let $d \in \mathbb{R}^n$ be a data sequence, and let $\mu(d)$ and $\sigma^2(d)$ be the (sample) mean and variance of the data d. Let p_d be the probability measure on $\mathbb{R}$ such that the density of p_d is a normal distribution with mean 0 and variance $\sigma^2(d)$. Let $\alpha \in (0, 1)$ be a fixed significance level, and define a method m such that m_d assigns (i) probability one to $\theta_{\mu \geq 0}$ if $p_d(x \in \mathbb{R} : x \geq \mu(d)\}) \geq \frac{1-\alpha}{\ln |d|}$ and (ii) probability one to $\theta_{\mu < 0}$ otherwise.

5.5 Convergence

Fix some natural number n. Given a method m and world w, define $p^n_{w,m}$ to be the unique measure on $\langle \Theta^n, \otimes_{k \leq n} \mathbb{B}(\Theta) \rangle$ satisfying the following. For all "rectangles" $E_1 \times E_2 \times \ldots \times E_n \in \otimes_{k \leq n} \mathbb{B}(\Theta)$ (i.e., $E_k \in \mathbb{B}(\Theta)$ for all $k \leq n$):

$$p^n_{w,m}(E_1 \times E_2 \times \ldots \times E_n) = \int_{D^n} \prod_{k \leq n} m_{\delta \restriction k}(E_k) \; dc^n_w(\delta)$$

where (1) c^n_w is the unique measure such that $c^n_w(F) = c_w(F \times D^{\mathbb{N}})$ for all $F \in \otimes_{k \leq n} \mathcal{D}$, and (2) $\delta \in D^n$. Under $p^n_{w,m}$, the probability of a method returning a sequence of answers is the chance of obtaining a data sequence δ (given by c^n_w) times the probability that the method returns a given answer (given by m) in response to δ.

It is easy to show that there is a unique probability measure $p_{w,m} \in \mathbb{P}(\otimes_{n \in \mathbb{N}})$ such that $p_{w,m}(E \times \Theta^{\mathbb{N}}) = p^n_{w,m}(E)$ for all $E \in \otimes_{k \leq n} \mathbb{B}(\Theta)$ and all $n \in \mathbb{N}$. Further, the following are events in $\otimes_{n \in \mathbb{N}} \mathbb{B}(\Theta) \rangle$, where $\theta \in \Theta$:

$$\begin{aligned} \{\overline{\theta}_n = \theta \text{ for large n}\} &:= \{\overline{\theta} \in \Theta^{\mathbb{N}} : (\exists n \in \mathbb{N})(\forall k \geq n)\overline{\theta}_k = \theta\} \\ \{\lim_{n \to \infty} \overline{\theta}_n = \theta\} &:= \{\overline{\theta} \in \Theta^{\mathbb{N}} : \lim_{n \to \infty} \overline{\theta}_n = \theta\} \end{aligned}$$

A method m is called **almost surely** (a.s.) convergent if $p_{w,m}(\lim_{n \to \infty} \overline{\theta}_n = \theta_w) = 1$ for all $w \in W$, and it is called **strongly almost surely** (s.a.s.) convergent if $p_{w,m}(\overline{\theta}_n = \theta_w$ for large $n) = 1$. When $\Theta = \{\{r\} : r \in \mathbb{R}^d\}$ is a parametric model, then a.s. convergence as defined here is the standard notion of a.s. convergence of a parameter estimator. It is trivial to show that, if Θ is finite, then a.s. convergence entails s.a.s. convergence.

Example: In the learning problem L_N, the method defined above is a.s. convergent by the second Borel Cantelli Lemma. Hence, it is s.a.s. convergent by the previous remark and the fact that Θ is finite.

5.5.1 Expert Networks

A **network** is a finite undirected graph G; vertices of G are called **agents**. A **group** is a set of agents $J \subseteq G$. For any $g \in G$, let $N_G(g) \subseteq G$ denote the group of agents $g' \in G$ such that g and g' are incident to a common edge. Call $N_G(g)$ the **neighborhood** of g, and call its elements **neighbors** of g. For simplicity, I assume every agent is her own neighbor. When G is clear from context, I will write $N(g)$ instead of $N_G(g)$.

An **expert network** $\mathcal{E}$ is a pair $\langle G, \langle L_g \rangle_{g \in G}, \langle m_g \rangle_{g \in G} \rangle$ such that G is a network, L_g is a learning problem for each agent $g \in G$, and m_g is a method for L_g. For all $g \in G$, let Q_g be the question confronted by g; define Θ_g, $c_{w,g}$, etc., similarly. An expert network can be represented by a colored undirected graph such that two vertices g and g' are the same color just in case $Q_g = Q_{g'}$.

Example: In the example in the body of the paper, the expert networks consist of agents confronted with instances of learning problem L_N. Note different agents may sample from different normal distributions.

Let $\Theta_{\mathcal{E}} = \{\Theta_g : g \in G\}$ be the set of questions faced by agents in the expert network $\mathcal{E}$, and let $A_{\mathcal{E}} = \times_{\Theta \in \Theta_{\mathcal{E}}} \Theta$ be the set of answers to all questions raised in the expert network. Define $\Theta_{\mathcal{E}-g} = \Theta_{\mathcal{E}} \setminus \{\Theta_g\}$ to be the set of questions faced by agents other than g, and $A_{\mathcal{E}-g} = \times_{\Theta \in \Theta_{\mathcal{E}-g}} \Theta$ be all possible answers.

For brevity, I introduce the following notation conventions. θ will represent an answer to a *single* question Θ, and a designates answers to *several* questions. Generally, a will be a member of $A_{\mathcal{E}}$ or of $A_{\mathcal{E}-g}$. The bolded letter $\boldsymbol{a}$ will indicate a group's answers to several questions; so $\boldsymbol{a} \in (A_{\mathcal{E}})^J$ or $\boldsymbol{a} \in (A_{\mathcal{E}-g})^J$ for some $J \subseteq G$. Finally, I use the "bar-notation" $\overline{\boldsymbol{a}}$ to indicate a sequence of group answers to several questions (so $\overline{\boldsymbol{a}} \in ((A_{\mathcal{E}})^J)^{<\mathbb{N}}$).

Recall, there is a Borel algebra $\mathbb{B}(\Theta)$ over each $\Theta \in \Theta_{\mathcal{E}}$. Hence, one can define $\mathcal{A}_{\mathcal{E}}$ be the product σ-algebra on $A_{\mathcal{E}} = \times_{\Theta \in \Theta_{\mathcal{E}}} \Theta$, and similarly for $\mathcal{A}_{\mathcal{E}-g}$. It is easy to check the following are events in these algebras:

$$\begin{aligned} \{(\forall g \in G)\overline{\boldsymbol{a}}_{n,g} = a \text{ for large } n\} &= \{\overline{\boldsymbol{a}} \in ((A_{\mathcal{E}})^G)^{\mathbb{N}} : (\exists n \in \mathbb{N})(\forall k \geq n)(\forall g)\overline{\boldsymbol{a}}_{n,g} = a\} \\ \{(\forall g \in G) \lim_{n\to\infty} \overline{\boldsymbol{a}}_{n,g} = a\} &= \{\overline{\boldsymbol{a}} \in ((A_{\mathcal{E}})^G)^{\mathbb{N}} : (\forall g \in G) \lim_{n\to\infty} \overline{\boldsymbol{a}}_{n,g} = a\} \end{aligned}$$

5.5.2 Testimonial Norms

A **testimonial norm** (TN) is a class of functions $\tau_{\mathcal{E},g} : ((A_{\mathcal{E}-g})^{N(g)})^{<\mathbb{N}} \to \mathbb{P}(\mathcal{A}_{\mathcal{E}-g})$, where $\mathcal{E}$ is an expert network and g is an agents in $\mathcal{E}$. Informally, a TN specifies a probability distribution over answers to questions outside g's specialty given what g's neighbors have reported in the past.

To define the six TNs introduced in the body of the paper, let $\overline{\boldsymbol{a}} \in ((A_{\mathcal{E}-g})^{N(g)})^{<\mathbb{N}}$ be an arbitrary sequence of answer reports of g's neighbors to all questions of interest. Suppose $\overline{\boldsymbol{a}}$ has length n. Recall that, for each agent h in g's neighborhood and each $\Theta \in \Theta_{\mathcal{E}-g}$, the symbol $\overline{\boldsymbol{a}}_{n,h,\Theta}$ represents h's report to question Θ on stage n (i.e. the last stage of a). Similarly, if $a \in A_{\mathcal{E}-g}$ is an answer to all questions outside of g's area of expertise, and if $\Theta \in \Theta_{\mathcal{E}-g}$ is one such question outside of g's area of expertise, then a_Θ is the answer a provides to the question Θ.

Example: Reidianism is the norm such that for all $a \in A_{\mathcal{E}-g}$:

$$\tau_{\mathcal{E},g}(\overline{\boldsymbol{a}})(a) = \prod_{\Theta \in \Theta_{\mathcal{E}-g}} \frac{|\{h \in N(g) : a_\Theta = \overline{\boldsymbol{a}}_{n,h,\Theta}\}|}{|N(g)|}$$

In other words, an answer θ to a given question Θ is chosen to be the proportion of one's neighbors that report θ on the most recent stage. Answers to different questions are chosen independently of one another, so the probability of choosing a sequence of answers a is the product of the probabilities of choosing each element a_Θ of the sequence.

To define e-trusting and proximitism, replace "$h \in N(g)$" in the above definition by the requirement that h is an expert neighbor (if one exists) or is most proximate to such an expert. The majoritarian versions of all three norms can be defined similarly.

A few properties of TNs will play a critical role in proofs. Let τ be a TN. Suppose that $(*)$ $\tau_{\mathcal{E},g}(\overline{\boldsymbol{a}}) = \tau_{\mathcal{E},g}(\overline{\boldsymbol{b}})$ for all expert networks $\mathcal{E}$, all agents g in $\mathcal{E}$, and all answer reports $\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}} \in ((A_{\mathcal{E}-g})^{N(g)})^{<\mathbb{N}}$ with identical last coordinates (i.e., $\overline{\boldsymbol{a}}_{|\overline{\boldsymbol{a}}|} = \overline{\boldsymbol{b}}_{|\overline{\boldsymbol{b}}|}$). Then τ is said to be **Markov**, as its behavior depends only upon the last element of an answer sequence. It is said to be Markov with memory t if $(*)$ holds for any sequences $\overline{\boldsymbol{a}}$ and $\overline{\boldsymbol{b}}$ for which the last t coordinates are identical.

Given $\Theta \in \Theta_{\mathcal{E}-g}$ and some $\theta \in \Theta$, define $E(\theta) = \{a \in A_{\mathcal{E}-g} : a_\Theta = \theta\}$. A Markov TN τ with memory t is said to be **stable** if for all $\overline{\boldsymbol{a}} \in ((A_{\mathcal{E}-g})^{N(g)})^{<\mathbb{N}}$, if $\overline{\boldsymbol{a}}_{k,h,\Theta} = \theta$ for all $h \in N(g)$ and all k such that $|a| - t \leq k \leq |a|$, then $\tau_{\mathcal{E},g}(\overline{\boldsymbol{a}})(E(\theta)) = 1$. Finally, a TN is said to be **sensitive** if for all expert networks $\mathcal{E}$, all agents g in the network, and all $\Theta \in \Theta_{\mathcal{E}}$, there is some $\epsilon > 0$ and some $J \subseteq PN(g, \Theta)$ such that for all $\overline{\boldsymbol{a}} \in ((A_{\mathcal{E}-g})^{N(g)})^{<\mathbb{N}}$, if $\overline{\boldsymbol{a}}_{|\overline{\boldsymbol{a}}|,h,\Theta} = \theta$ for all $h \in J$, then $\tau_{\mathcal{E},g}(\overline{\boldsymbol{a}})(E(\theta)) > \epsilon$. By construction, the six TNs in the body of the paper are Markov, and stable. All said TNs except majoritarian Reidianism and majoritarian e-trusting are also sensitive.

A **group testimonial norm** (or GTN for short) is a proper class function from expert networks to vectors of TNs for each agent in the network. A GTN is called **pure** if it is a constant function; it is called **mixed** otherwise.

5.6 Scientific Communities, Probabilities over Answer Sequences, and More on Convergence

A **scientific network** is a pair $S = \langle \mathcal{E}, \langle \tau(g) \rangle_{g \in G} \rangle$ consisting of an expert network $\mathcal{E}$ and an assignment of TNs $\tau(g)$ to each agent g in the network. $\tau(g)_{\mathcal{E},g}$ is abbreviated by $\tau_{\mathcal{E},g}$ below, as no confusion will arise.

Define $W_{\mathcal{E}} = \{W_g : g \in G\}$, and let $\overline{w} \in \times_{W \in W_{\mathcal{E}}} W$ be the true state of the world for all questions faced by agents in the network. Recall, in a given world, an agent's methods induces a probability measure over answer sequences *within* her area of expertise. Moreover, GTNs specify the probability that agents will assign to answers *outside* their respective specialties. Therefore, given a scientific network S and world $\overline{w} \in \times_{W \in W_{\mathcal{E}}} W$, one can define a probability measure $p_{\overline{w},S}$ over infinite sequences of answers for the entire network $((A_{\mathcal{E}})^G)^{\mathbb{N}}$ (where, the events are those in the product algebra). Defining $p_{\overline{w},S}$ is straightforward but tedious. So details are omitted.

Say an expert network $\mathcal{E}$ is **s.a.s methodologically convergent** if the methods employed by each agent are s.a.s. Given $\overline{w} \in \times_{W \in W_{\mathcal{E}}} W$, let $a(\overline{w}) \in A_{\mathcal{E}}$ be the unique answer sequence such that $\overline{w} \in a(\overline{w})$. That is, $a(\overline{w})$ is the sequence of true answers to every question if $\overline{w}$ describes the true state of the world. Say a GTN is **s.a.s testimonially convergent** if for all scientific networks $S = \langle \mathcal{E}, \langle \tau(g) \rangle_{g \in G} \rangle$:

$$p_{\overline{w},S}((\forall g \in G)\overline{\boldsymbol{a}}_{n,g} = a(\overline{w}) \text{ for large } n) = 1$$

whenever $\mathcal{E}$ is a connected, s.a.s. methodologically convergent network.

5.7 Proofs of Theorems

Given an expert network $\mathcal{E}$ and $\overline{w} \in \times_{W \in W_{\mathcal{E}}} W$, define $\boldsymbol{a}(\overline{w}) \in (A_{\mathcal{E}})^G$ to be the vector representing the state in which all agents believe $a(\overline{w})$. Let $\boldsymbol{A}(\overline{w}) = \{\boldsymbol{a} \in (A_{\mathcal{E}})^G : w_g \in \boldsymbol{a}_{g,\Theta_g}\}$ be the set of belief vectors in which every agent holds a true belief *in her own specialty.* Next, define by recursion:

$$\begin{aligned} E_0(\overline{w}) &= \{\overline{\boldsymbol{a}} \in ((A_{\mathcal{E}})^G)^{\mathbb{N}} : (\forall n \in \mathbb{N})\overline{\boldsymbol{a}}_n \in \boldsymbol{A}(\overline{w})\} \\ E_{n+1}(\overline{w}) &= \{\overline{\boldsymbol{a}} \in ((A_{\mathcal{E}})^G)^{\mathbb{N}} : (\forall k \geq n+1)\overline{\boldsymbol{a}}_k \in \boldsymbol{A}(\overline{w})\} \setminus E_n(\overline{w}) \end{aligned}$$

So $E_n(\overline{w})$ is the event that n is the first stage at which every agent has converged to the true answer in her specialty. Finally, let $X_n : (A_{\mathcal{E}}^G)^{\mathbb{N}} \to A_{\mathcal{E}}^G$ be the function $\overline{\boldsymbol{a}} \mapsto \overline{\boldsymbol{a}}_n$ that represents the beliefs of *all* agents, to *all* questions on stage n. In the following lemma, let $S = \langle \mathcal{E}, \langle \tau(g) \rangle_{g \in G} \rangle$ be a scientific network and $\overline{w} \in \times_{W \in W_{\mathcal{E}}} W$. Suppose that $\mathcal{E}$ is methodologically s.a.s. convergent and that $\tau(g)$ is Markov for all $g \in G$.

Lemma 1 *Then $\langle X_n, E_n(\overline{w}) \rangle_{n \in \mathbb{N}}$ is a uniform pc-Markov process over the state space $\boldsymbol{A}(\overline{w})$ with respect to $p_{\overline{w},S}$.*

Proof: Since $\mathcal{E}$ is s.a.s. methodologically convergent, it follows that $p_{\overline{w},S}(\cup_{n \in \mathbb{N}} E_n(\overline{w})) = 1$. The events $E_n(\overline{w})$'s are disjoint by construction.

Conditional on E_n, each agent's beliefs change only *outside* her specialty at every stage $k \geq n+1$. Hence, agents' beliefs at any stage $k \geq n$ depend only upon TNs and not upon data. Since the TNs are Markov, the vectors of all agents beliefs at stages past n, represented by $\langle X_k \rangle_{k \geq n}$, form a time-homogeneous, Markov process conditional on E_n as desired.$\dashv$

The next theorem corresponds to Theorem 1 in the text.

Theorem 6 *Suppose that, for all $g \in G$, the norm $\tau(g)$ is also stable and sensitive. Then $\langle X_n, E_n(\overline{w}) \rangle_{n \in \mathbb{N}}$ is a uniformly absorbing pc-Markov process with respect to $p_{\overline{w},S}$, where the unique absorbing state is $\boldsymbol{a}(\overline{w})$.*

Proof: By the previous lemma, $\langle X_n, E_n(\overline{w}) \rangle_{n \in \mathbb{N}}$ is a uniform pc-Markov process. For all agents g, one can use sensitivity to show, by induction on g's length n from a Θ-expert, that there is some non-zero probability that g will believe an answer θ to Θ exactly n many stages after all the most proximate Θ experts to g believe θ. Again, using stability and induction, one can show that, for all natural numbers n and k, if g believes θ on stage n and all the most proximate Θ experts to g continue to believe θ for k stages, then g will believe θ on stage $n+k$. Since the network is s.a.s. convergent, this suffices to show that true beliefs will eventually propagate through the entire network. By the stability and Markov property of the TNs, the network will be absorbed in this state.$\dashv$

5.8 Modeling Miscommunication

For the following theorems, suppose each agent's question has only two answers. In order to model miscommunication, one needs only to alter the definition of the measure $p_{\overline{w},S}$, so that, on each stage, for all her neighbors, an agent reports the answer other than the one she believes with some fixed probability $\epsilon > 0$. Call the measure induced by this process $p_{\overline{w},S,\epsilon}$. Below, retain the assumptions of the previous theorems and add the further assumption that, for all $g \in G$, the function $\tau(g)$ is one of the six TNs from the body of the paper.

Theorem 7 *Then $\langle X_n, E_n(\overline{w}) \rangle_{n \in \mathbb{N}}$ is a uniformly regular pc-Markov process (over state space $\boldsymbol{A}(\overline{w})$) with respect to $p_{\overline{w},S,\epsilon}$.*

Proof: By the same reasoning as above, the process is a pc-Markov process. So it suffices to show it is regular. In fact, since each of the six TNs has a memory of length one, the process can transition from any state in $\boldsymbol{A}(\overline{w})$ to another in exactly one step. To show this, I show that any agent's belief, with respect to any question, changes with positive probability and stays the same with positive probability. This suffices because there are only two answers to a question.

Consider a fixed agent g and a fixed question Q. Since there are two possible answers, the agent's belief with respect to Q can be represented by a 0 or 1, and her neighbors beliefs with respect to Q can be represented by a binary vector $\boldsymbol{a}$. Now each of the six TNs has the following property: there are binary vectors $\boldsymbol{b}_{stay}$ and $\boldsymbol{b}_{change}$ such that, (i) if g thinks her neighbors' beliefs with respect to Q are represented by $\boldsymbol{b}_{stay}$, then g's beliefs with respect to Q will remain the same with positive probability, and (ii) if g believes her neighbors' beliefs are represented by $\boldsymbol{b}_{change}$, then g's beliefs will change with positive probability. For instance, if g is a Reidian who currently believes 0, then the constant vector containing only zeros is one example that could be $\boldsymbol{b}_{stay}$, and the constant vector containing only ones is one example of $\boldsymbol{b}_{change}$.

Let $\boldsymbol{a}$ represent g's neighbors' current beliefs, and Let n be the number of entries in the vector differs from the vector $\boldsymbol{b}_{stay}$. Then, by the definition of miscommunication, the probability that g will think her neighbors believe $\boldsymbol{b}_{stay}$ is ϵ^n. By definition of $\boldsymbol{b}_{stay}$, if g believes her neighbors believe $\boldsymbol{b}_{stay}$, then g will retain her belief with some positive probability δ. So the probability that g's belief will stay the same is at least $\delta \cdot \epsilon^n$, which is positive. A similar argument shows that g's belief changes with respect to question Q with positive probability. ⊣

The next theorem corresponds to Theorems 3 and 2 in the body of the paper.

Theorem 8 *The error rate approaches a fixed positive number.*

Proof: By the previous lemma and Theorem 5, there is a limiting probability distribution over states (i.e. specifications of beliefs for every agent in the network) of the PC-Markov process describing the evolution of agents' beliefs, and that distribution does not depend on agents' initial beliefs. In each such state, agents have some non-negative number of erroneous beliefs. The error rate is the expectation of error relative to this limiting probability distribution over states. Note every state has some positive probability by the argument in the previous theorem. So there is some non-zero probability of having erroneous beliefs, which entails the error rate is positive.⊣