Knowledge in a Social Network
PREPRINT

Staffan Angere
Lund University
staffan.angere@fil.lu.se

March 29, 2010

Abstract

The purpose of this paper is to present a formal model of social networks suitable for studying questions in social epistemology. We show how to use this model, in conjunction with a computer program for simulating groups of inquirers, to draw conclusions about the epistemological properties of different social practices. This furnishes us with the beginnings of a systematic research program in social epistemology, from which to approach problems pertaining to epistemic value, optimal organization, and the dynamics of multi-agent inquiry.

1 The Epistemology of Social Networks

Social epistemology concerns the pursuit of truth in groups of inquirers. While individualistic epistemology studies what properties of an individual are conducive to the attainment of truth, social epistemology centers on those properties of groups that are irreducible (or at least very hard to reduce) to properties of these groups' members, such as the communicative practices and the division of interrogative labor. But one may also hold that the goals differ: while individualistic epistemology studies knowledge in a single person, social epistemology is concerned with knowledge in society.

Given how difficult the search for an exact definition of knowledge in an individual has been, the question of when a society has knowledge may appear insurmountable. However, this does not have to stop us from using definitions that are good enough for most purposes, and we will study some of these later on. It may even be that when we get clear on what we want social epistemology to deliver, the question of whether that thing is to be called 'social knowledge' properly can take a back seat.

For the purposes of this paper, I will assume that social epistemology aims at maximizing the strength of people's true beliefs in a society. Now, true belief is a very weak form of knowledge (Alvin Goldman, in his seminal book about social epistemology [5, p. 23], calls it 'W-knowledge'), but it is clearly something that

is epistemically valuable. It also has the advantage of being subject to formal modelling, and there are good formal theories of truth and belief around. In contrast to this, the question of when a belief is justified seems considerably harder. Since our aim here is to apply formal methods to questions in social epistemology, it seems prudent to bypass these issues for now.

Starting our model-building, we define a *social network* as a group of people, or people-like entities (such as some organizations), among which there are practices regarding communication. Some examples are university departments, peer-review boards, circles of friends, or even entire societies. As a limiting case, the entirety of humanity during a period can be seen as such a network. The 'during a period' clause is necessary here: we are taking the network, i.e. the communicative and investigative practices themselves, to be fairly stable. We also assume that what participants are in the network is unchanging.

Formally, we can take a social network S to be a set Γ, which we for reasons that will be made obvious shortly will call the set of *inquirers*, together with a binary relation R on Γ, which we call the *network structure*. This means that a social network is a directed graph. From another viewpoint, it can thus be taken to be a Kripke model for a normal modal logic, and we can say that $\alpha \vDash p$ iff α believes p, and introduce an operator $\Box p$ to say that *every source α has available believes that p* or possibly *every source α has available says that p*. This way, we can extend logics of knowledge, belief and belief revision into the social world. Such approaches are explored in the fields of Public Announcement Logic (cf. [1]) and Dynamic Epistemic Logic (cf. [3]).

But there is another paradigm of contemporary formal epistemology as well: Bayesianism. Here, the epistemic state of a person α at time t is assumed to be given by a *credence function* $C_\alpha^t : \mathcal{L} \to [0, 1]$ instead of just a set of sentences. $\mathcal{L}$ can be taken to be a classical propositional language, and C_α^t is assumed to fulfil the standard axioms of a probability measure. Due to its close connection to probability theory, Bayesianism is well suited for statistical models, and as the models that we will investigate in this paper are of this kind, we will adopt the Bayesian approach.

In a Bayesian model, there is no all-or-nothing concept of belief, and we should also wish to say that there is no all-or-nothing concept of knowledge either, even if we interpret it in the weak sense that Goldman does. However, the important point here is not the knowledge concept itself, but how we evaluate epistemic states. Furthermore, the Bayesian concept does admit comparisons (a kind of "degree of knowledge"), and this is all we need for such evaluation. This approach will be explored in section 3 below.

The inquirers of a social network naturally have an internal structure, and it is necessary for us to understand this structure, in order to be able to evaluate the network's epistemological properties. For the purposes of this paper, let us confine ourselves to the case where inquiry is aimed at discovering whether a single proposition p is true or false. We assume, conventionally, that p happens to be true, since this will simplify calculations further on. At the very least, every inquirer will then have a credence $C_\alpha^t(p)$ in p, which is a real number between 0 and 1, for every moment t. But in order to investigate epistemological behavior,

we need more than this. We also have to model the way that the participants receive new information. There are two fundamentally different inlets for this: inquiry and communication.

Inquiry can here be taken to include any kind of method of altering a credence function which does not base itself on information given by anyone else in the network. The paradigmatic cases of inquiry are observation, experiment, and perhaps taking advice from persons outside S. In order for our model to be applicable, the opinions of participants in S must have fairly little effect on such "external" persons' opinions, however. An example where this holds is where the communication is effected through a book, as when a modern philosopher reads Kant. Kant himself does not need to be taken as a part of our social network, since his opinions (perhaps unfortunately!) cannot be affected by anything we do.

Not all paritcipants' approaches to inquiry are the same, and they tend to vary both in their degree of activity and their effectiveness. Let $S^t_{\iota\alpha}p$ be the proposition 'α's inquiry gives the result that p at time t', $S^t_{\iota\alpha}\neg p$ be the proposition 'α's inquiry gives the result that not-p at t', and $S^t_{\iota\alpha} =_{df.} S^t_{\iota\alpha}p \vee S^t_{\iota\alpha}\neg p$ the proposition that α's inquiry gives some result at t. We represent the participants' properties *qua* inquirers by two probabilities: the chance $P(S^t_{\iota\alpha})$ that, at any moment t, α receives a result from her inquiries, and the chance $P(S^t_{\iota\alpha}p \mid S^t_{\iota\alpha} \wedge p)$ that, when such a result is obtained, it is the right one.[1] $P(S^t_{\iota\alpha})$ will be referred to as α's *activity*, and $P(S^t_{\iota\alpha}p \mid S^t_{\iota\alpha} \wedge p)$ as her *aptitude*. An inquirer without interest in p would generally have a low value of $P(S_{\iota\alpha})$, while one very interested in p, but engaging in inquiry using faulty methods would have a high value of $P(S_{\iota\alpha})$, but an aptitude close to 0.5, or even below that. In the latter case, the results of her inquiry would actually be negatively correlated with the truth. As a simplification, we will assume α's activity and aptitude to be constant over time, so we will generally write them without the time index t.

Just as inquiry represents the flow of information into the network, communication deals with how this information is disseminated. While, as a first approximation, we may be interested only in whether or not α can receive information from β, we generally need to go deeper. Thus, we take the network R to be a set of what we will call *links*, each corresponding to a communication channel. Such channels can be as direct as conversation, or mediated like a blog or an instant messaging system. Generally, however, we may want to say that if the "mediator" is able to choose which messages to transmit, it may be more apt to be represented as an inquirer instead of a link, even if it engages in no inquiry itself (this is one case where we may want to let $P(S_{\iota\alpha}) = 0$). A scientific journal could be an example of this. The links from other inquirers to the journal then summarize the peer review process, and the links from the journal to the other inquirers represent their individual practices of reading that

[1] This is actually only half the story: $P(S^t_{\iota\alpha}p \mid S^t_{\iota\alpha} \wedge p)$ gives the chance of a true result when the result indicates that p is the case, and does not say anything about reliability when $\neg p$ is reported. To simplify matters, we assume that these probabilities are equal, i.e. that the chance that inquiry gives a true result does not depend on whether p is true or false.

journal.

Like inquirers, links have an internal structure. Although they do not have anything corresponding to an aptitude, since the messages they transmit are chosen by the senders and thus not subject to random variation, they have degrees of activity. Two inquirers who seldom talk have weak links between them, while researchers who sit in the same room may be assumed to have stronger links. Analogously to the inquiry notation we define

$$\begin{aligned} S^t_{\beta\alpha}p &=_{df.} \beta \text{ says that } p \text{ to } \alpha \text{ at } t \\ S^t_{\beta\alpha}\neg p &=_{df.} \beta \text{ says that not-}p \text{ to } \alpha \text{ at } t \\ S^t_{\beta\alpha} &=_{df.} \beta \text{ says either that } p \text{ or that not-}p \text{ to } \alpha \text{ at } t \end{aligned}$$

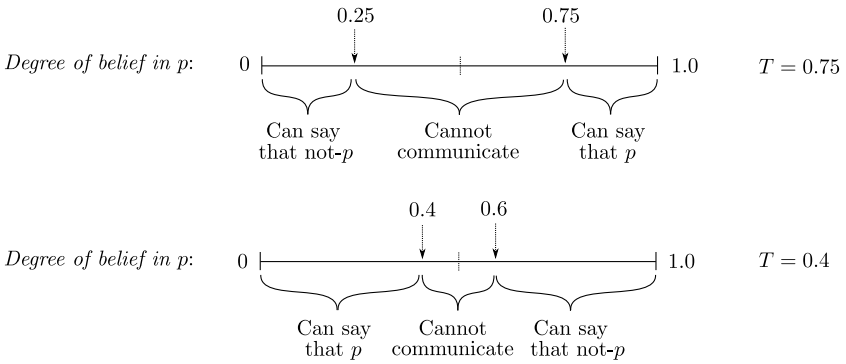
This strength of a link $\beta\alpha$ is then representable as a probability $P(S_{\beta\alpha})$, being the chance that β communicates that p or that not-p to α, at any given moment t.

Given that β *does* communicate with α, what does she say? *Prima facie*, it seems that β somehow would indicate how strongly she believes that p, i.e. she would communicate her credence. But this credence is not in general available to her. As a subjective degree of belief, it is a product of betting behavior, and it takes experiments to determine it rather than introspection. We may of course represent β's beliefs about her degree of belief as credences as well, but this only pushes the problem further back: somehow, we will still need to determine exactly what to say, given a credence distribution over possible degrees of belief.

Instead, we will adopt a simplified model, but one which hopefully still has enough complexity to be interesting. We assume that unless she is out to mislead, β will say p if she believes p, and not-p if she believes not-p. But belief is not an all-or-nothing matter in the Bayesian tradition, so how sure must she be of p to say it? And, conversely, if she *is* out to mislead, how sure must she be of p to say that not-p? These questions are answered by a property of the link $\beta\alpha$ that we will call its *threshold*: a value $T_{\beta\alpha}$ between 0 and 1, such that

- if $T_{\beta\alpha} > 0.5$, β tells α that p only if $C_\beta(p) \geqslant T_{\beta\alpha}$, and that not-$p$ only if $C_\beta(p) \leqslant 1 - T$, and
- if $T_{\beta\alpha} < 0.5$, β tells α that p only if $C_\beta(p) \leqslant T_{\beta\alpha}$, and that not-$p$ only if $C_\beta(p) \geqslant 1 - T$.
- if $T_{\beta\alpha} = 0.5$, β can tell α that p or that not-p independently of what she believes. We model this by letting her pick which to say randomly.

Thus, a link that consists of systematic lying will have a threshold below 0.5, and one that proceeds through truth-telling (at least insofar as the link's source is aware) has a threshold above 0.5. The effect of $T_{\beta\alpha}$ on when the link $\beta\alpha$ is permissible for communication by β is illustrated in the below image:



Are there communicative practices not representable as links of this kind? We have already mentioned the scientific journal, which may be more reasonable to model as an inquirer. But what of mass broadcasting systems, like a TV station, or the above-mentioned blog? A link, as we have defined it, has a single source and a single recipient. However, nothing stops us from representing a broadcasting system by a set of links — one for each viewer. If we interpret $P(S_{\beta\alpha})$ as the probability that α actually *hears* what β says, rather than just the probability that β says something, then even a broadcasting system ought to have different links for different viewers.

2 Trusting oneself and others

We have described how the participants in a social network engage in inquiry and communicate, but we have as yet said nothing about how they react to the results of these practices. It seems that, in general, hearing that p from someone, or receiving a result indicating p from inquiry, should influence an inquirer's credence in p. Fortunately, Bayesianism has a universal answer to the question of how this should be done: belief update proceeds through conditionalization.

But conditionalizing on p whenever one hears that p is not reasonable. Straight conditionalization really works only for infallible sources, and in general an inquirer only takes messages proclaiming p as indications of its truth, and not as conclusive verification. One solution is to use Jeffrey conditionalization instead of regular conditionalization [7]. This is more of a promissory note than a solution as such, since we still need to decide on a degree of belief to set p to, when one hears that p. Furthermore, inquirers can receive several messages at the same time, and some of these may even contradict one another, so what we need is a framework that allows us to handle such cases.

One way forward is as follows. To start with, it seems admissible for α to treat both her own inquiry and the things said to her in roughly the same way: as indications of whether or not p is the case. We refer to α's inquiry ι and the other inquirers $\beta, \gamma, \ldots$ who can talk to her as her *sources*. Each of these may have different connections to the truth, and we can represent these by the probability of a source to give the right answer. More specifically, we define α's

source σ's *reliability* as

$$R_{\sigma\alpha} =_{df.} P(S_{\sigma\alpha}p \mid S_{\sigma\alpha} \wedge p) = P(S_{\sigma\alpha}\neg p \mid S_{\sigma\alpha} \wedge \neg p)$$

This definition presupposes that the probability that any source gives the answer p if p is the case is to be equal to the probability that it gives the answer not-p, if not-p is the case. This *source symmetry* simplifies our calculations, although it can be relaxed if we encounter cases where it does not provide a reasonable approximation.[2]

It follows at once that the reliability of α's inquiry is identical to her aptitude. For other sources, it is an abstraction based on those sources' performances as indications of truth. In general, an inquirer has no direct access to this value, but this does not stop her from forming beliefs about it. Since the number of possible values for the chance $R_{\sigma\alpha}$ is infinite, we need to represent α's credence as a density function instead of a regular probability distribution. Thus, for each inquirer α, each source σ, and each time t, we define a function $\tau^t_{\sigma\alpha} : [0,1] \to [0,1]$, called *$\alpha$'s trust function for σ at t*, such that

$$C^t_\alpha(a \leqslant R_{\sigma\alpha} \leqslant b) = \int_a^b \tau^t_{\sigma\alpha}(\rho)\, d\rho$$

for a, b in $[0,1]$. This function is uniquely defined up to a set of measure 0, according to the Radon-Nikodym theorem. $\tau_{\sigma\alpha}(\rho)$ then gives the credence density at ρ, and we can obtain the actual credence that α has in propositions about the reliability of her sources by integrating this function. We will also have use for the expression $1 - \tau^t_{\sigma\alpha}$ (which represents α's credence density for propositions about σ *not* being reliable) and we will refer to this function as $\overline{\tau}^t_{\sigma\alpha}$.

Now, it is obvious that an inquirer's credences about chances should influence her credences about the outcomes of these chances. The way this should be done is generally known under a name Lewis gave to it: the principal principle [12]. This says that if α knows that the chance that an event e will happen is ρ, then her credence in e should be exactly ρ. Applied to our case, this means that the following principle (PP) must hold:

$$C^t_\alpha(S^t_{\sigma\alpha}p \mid S^t_{\sigma\alpha} \ \wedge \ R_{\sigma\alpha} = \rho \ \wedge \ p) = \rho$$
$$C^t_\alpha(S^t_{\sigma\alpha}\neg p \mid S^t_{\sigma\alpha} \ \wedge \ R_{\sigma\alpha} = \rho \ \wedge \ \neg p) = \rho$$

for all t, i.e. α's credence in σ giving the report p, given that the source gives any report at all, that σ's reliability is ρ, and that p actually is the case, should be ρ.

We also have use for an independence postulate. While not strictly necessary, such a postulate will simplify calculations and modelling considerably. The

[2] There is another way to define reliability as well, according to which it is the probability that p is the case, given that p is reported. They are, however, interderivable, and as the definition used here is simpler for our purposes, we have adopted it. The difference corresponds to Goldman's distinction between *reliability* and *power*. Cf. [4].

independence assumption we use here will be referred to as *communication independence* (CI):

$$C^t_\alpha\left(p \wedge S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho\right) = C^t_\alpha(p)\; C^t_\alpha(S^t_{\sigma\alpha})\; R^t_{\sigma\alpha}(p)$$

Communication independence implies that whether σ says anything is independent of whether or not p actually is true, as well as of what reliability σ has. This is true in the current model, since we have assumed that a source's reliability for reporting that p, given that p is the case, is the same as its reliability for reporting that $\neg p$ is the case, given that $\neg p$.

From (PP) and (CI) we can derive the following expression for α's credence in σ's reliability (see the appendix for the actual derivation):

$$C^t_\alpha(S^t_{\sigma\alpha}p \mid p) = C^t_\alpha(S^t_{\sigma\alpha}) \int_0^1 \rho\; \tau^t_{\sigma\alpha}(\rho)d\rho$$

But the right-hand integral is the also *expected value* $\langle\tau^t_{\sigma\alpha}\rangle$ of the trust function $\tau^t_{\sigma\alpha}$, so we can rewrite this as

$$(*)C^t_\alpha(S^t_{\sigma\alpha}p \mid p) = C^t_\alpha(S^t_{\sigma\alpha})\langle\tau^t_{\sigma\alpha}\rangle$$

Using this equation, an application of Bayes' theorem together with the theorem of total probability gives us that

$$C^t_\alpha(p \mid S^t_{\sigma\alpha}p) = \frac{C^t_\alpha(p)\langle\tau^t_{\sigma\alpha}\rangle}{C^t_\alpha(p)\langle\tau^t_{\sigma\alpha}\rangle + C^t_\alpha(\neg p)\langle\overline{\tau}^t_{\sigma\alpha}\rangle}$$
$$C^t_\alpha(p \mid S^t_{\sigma\alpha}\neg p) = \frac{C^t_\alpha(p)\langle\overline{\tau}^t_{\sigma\alpha}\rangle}{C^t_\alpha(p)\langle\overline{\tau}^t_{\sigma\alpha}\rangle + C^t_\alpha(\neg p)\langle\tau^t_{\sigma\alpha}\rangle}$$

Since we, by the requirement of conditionalization, must have that $C^{t+1}_\alpha(p) = C^t_\alpha(p \mid S_{\sigma\alpha}p)$ whenever σ is the only source giving information to α at t, this formula completely determines how α should update her credences in such a case. Some of the consequences of this can be summarized qualitatively. We say that σ is *trusted* when $\langle\tau^t_{\sigma\alpha}\rangle > 0.5$, *distrusted* when $\langle\tau^t_{\sigma\alpha}\rangle < 0.5$ and *neither trusted nor distrusted* when $\langle\tau^t_{\sigma\alpha}\rangle = 0.5$. Furthermore, we say that a message m is *expected* if $C^t_\alpha(p) > 0.5$ and $m = p$ or $C^t_\alpha(p) < 0.5$ and $m = \neg p$, *unexpected* if $C^t_\alpha(p) > 0.5$ and $m = \neg p$ or $C^t_\alpha(p) < 0.5$ and $m = p$, and *neither expected nor unexpected* otherwise (i.e. if $C^t_\alpha(p) = 0.5$). The following table (whose derivation is left as an excercise for the reader) gives the qualitative rules for how belief is updated. A '+' means that the message reinforces α's current belief (i.e. her credence increases if above 0.5 and decreases if below 0.5), '-' that the strength of her belief is weakened (i.e. that her credence increases if below 0.5 and decreases if above 0.5), and '0' that her credence is unchanged.

Is source trusted?	Is message expected? Yes	Neither	No
Yes	+	0	-
Neither	0	0	0
No	-	0	+

The calculations become slightly more complex when we take into account the possibility of receiving several messages at the same time. Let Σ_α^t be the set of sources from which α receives information at t, and let $m_{\sigma\alpha}^t$ be the message (i.e. either p or not-p) that σ gives α at t. Conditionalization requires that

$$C_\alpha^{t+1}(p) = C_\alpha^t\left(p \;\middle|\; \bigwedge S_{\sigma\alpha}^t m_{\sigma\alpha}^t\right) \tag{1}$$

where the conjunction runs over all sources σ in Σ_α^t. This may be a very complex expression, but we can simplify it if we take inquirers to treat their sources as independent, given the truth or falsity of p. Formally, this means that we adopt the following axiom of *source independence* (SI):

$$C_\alpha^t\left(\bigwedge S_{\sigma\alpha}^t m_{\sigma\alpha}^t \;\middle|\; p\right) = \prod C_\alpha^t\left(S_{\sigma\alpha}^t m_{\sigma\alpha}^t \mid p\right) \tag{2}$$

Is this a reasonable assumption to make? In certain cases, it may very well be. We generally tend to see our sources as not engaged in some kind of conspiracy so long as we do not receive positive indication that they are. It is this that makes us attach a greater degree of belief to propositions the more sources we hear them from. Although this is a simplification — there may be all kinds of underlying dependences in effect — it is a simplification that is necessary, not only for the epistemologist, but for everyday practice as well. It is reasonable as a default assumption.

Given source independence, the properties of individual links and inquirers determine how inquirers should update their beliefs when they receive new information. An application of Bayes' theorem together with the theorem of total probability gives that

$$C_\alpha^t\left(p \;\middle|\; \bigwedge S_{\sigma\alpha}^t m_{\sigma\alpha}^t\right) = \frac{C_\alpha(p)\prod C_\alpha^t\left(S_\sigma^t m_{\sigma\alpha}^t \mid p\right)}{C_\alpha(p)\prod C_\alpha^t\left(S_\sigma^t m_{\sigma\alpha}^t \mid p\right) + C_\alpha(\neg p)\prod C_\alpha^t\left(S_\sigma^t m_{\sigma\alpha}^t \mid \neg p\right)}$$

Since the values $C_\alpha^t\left(S_\sigma^t m_{\sigma\alpha}^t \mid p\right)$ and $C_\alpha^t\left(S_\sigma^t m_{\sigma\alpha}^t \mid \neg p\right)$ are determined by eq. $(*)$, this equation lets us infer exactly what degree of belief an inquirer α should give p, given the information she receives at any time.

Not only α's credence in p should be updated, however. Equally important is for α to keep track of how much to trust her sources. A source that generally gives very unlikely reports is unlikely to be veridical, and an inquirer should adjust her trust function in light of this. It turns out that our model already determines how to do this: given that α's trust function for the source σ is $\tau_{\sigma\alpha}^t$

at t, and that she receives the result that p, or that not-p from σ then, her new trust function $\tau_{\sigma\alpha}^{t+1}$ is given either by[3]

$$\tau_{\sigma\alpha}^{t+1}(\rho) = \tau_{\sigma\alpha}^{t}(\rho)\frac{\rho\, C_\alpha^t(p) + (1-\rho)C_\alpha^t(\neg p)}{\langle\tau_{\sigma\alpha}^t\rangle\, C_\alpha^t(p) + \langle\overline{\tau}_{\sigma\alpha}^t\rangle\, C_\alpha^t(\neg p)}$$

or by

$$\tau_{\sigma\alpha}^{t+1}(\rho) = \tau_{\sigma\alpha}^{t}(\rho)\frac{\rho\, C_\alpha^t(\neg p) + (1-\rho)C_\alpha^t(p)}{\langle\tau_{\sigma\alpha}^t\rangle\, C_\alpha^t(\neg p) + \langle\overline{\tau}_{\sigma\alpha}^t\rangle\, C_\alpha^t(p)}$$

depending on whether the report received from σ claims that p or that not-p.

That both C_α and $\tau_{\sigma\alpha}$ change as a result of inquiry and communication gives rise to complex interactions. Assume, for example, that α starts out with a credence in p slightly below 0.14, and some trust in her own inquiry, as indicated by a value of expected trust at 0.67. Assume that her inquiry keeps indicating that p actually is the case. How her trust in inquiry changes after t such inputs then depends crucially on the trust function τ she starts out with. But even if the *expected* trust is the same, the shape of the trust function influences the long-term behavior. Figure 1 illustrates three such scenarios, during a time lapse of 50 inputs. The table that follows gives the corresponding degrees of belief and values of expected trust for these.

	(a)		(b)		(c)	
Time	$C(p)$	$\langle\tau\rangle$	$C(p)$	$\langle\tau\rangle$	$C(p)$	$\langle\tau\rangle$
0	0.14	0.67	0.14	0.67	0.14	0.67
1	0.24	0.56	0.24	0.61	0.24	0.62
5	0.18	0.36	0.45	0.52	0.59	0.58
10	0.0	0.19	0.5	0.5	0.94	0.67
50	0.0	0.04	0.5	0.5	1.0	0.93

Here, we have three different ways that the inquirer's credence might evolve, given the same initial credence, the same obtained evidence, and the same initial expected trust. Although our inquirer happens to be a *perfect* inquirer insofar as her inquiry always gives the right result, the fairly low stability of her faith in inquiry in (a), together with her prior judgment that p is unlikely, conspire to make her distrust her own inquiry. This, in turn, gives rise to a vicious circle in which she becomes more and more convinced that p is false, and that her inquiry is negatively correlated with the truth.

In (b), the inquirer's trust happens to be just enough to counter her prior disbelief in p, but not enough to get her to believe that p, with the result that her credence converges to 0.5. In (c), her trust in inquiry is stable enough to overcome her prior disbelief, with the happy consequence that she converges on the truth.

What can we learn from these scenarios? The most important lesson is that trust is a complex issue, impossible to capture in a single number. Although

[3]See appendix 2 for the derivation

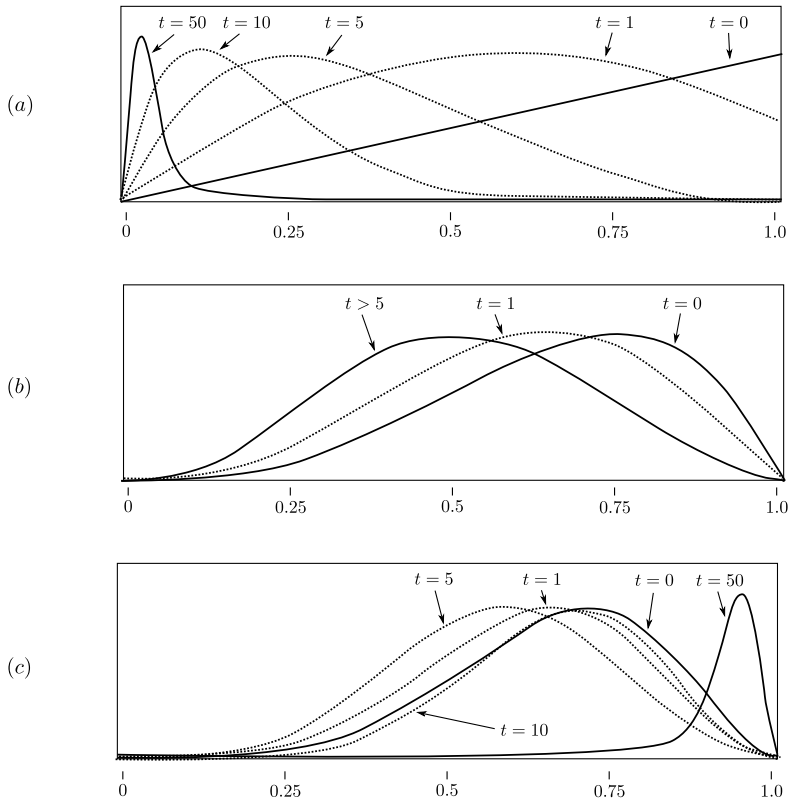


Figure 1: Influence of the shape of τ on α's credence in p.

I mentioned 'stability' in the preceding paragraphs, it is also the case that *two* numbers are insufficient, so stability should not be seen as some kind of complement to the expected trust. Although only the expected trust influences an inquirer's posterior credence for a single update, this does not hold for further updates. The general principle, statistically speaking, is that the result of n updates requires information about the first to nth central moments of τ. In the infinite limit, all of τ is necessary.

3 Truth, Error and Epistemic Value

In the preceding sections, we have presented a way to model social networks. There are several things we can ask about such a network, and the most important ones for the epistemologist concern how effective its participants are at getting to the truth of important questions. In other words, how good is the network, epistemologically speaking?

This question permits a multitude of answers, depending on how we explicate the notion of epistemc value. This subproblem of the general axiological problem

(how do we determine how good a certain state of the world is?) inherits its difficulties, although we will try to avoid these here as far as possible. An influential interpretation of the concept is given by Goldman [5, ch. 3], who equates it with *veritistic value*, which in turn is defined as the *average credence in the truth* in the society. In our terms, the veritistic value V of a social network state $S^t = \langle \Gamma^t, R^t \rangle$ at time t can be written as

$$V(S^t) = \frac{1}{|\Gamma^t|} \sum_{\alpha \in \Gamma^t} C^t_\alpha(p)$$

There is certainly something right about this formula: in general, it seems that if S^t_1 and $S^{t'}_2$ are states of the same social network, and, no one in $S^{t'}_2$ believes in p less strongly than she does in $S^{t'}_1$, then $S^{t'}_2$ should be at least as epistemically good as S^t_1. Thus, epistemic value should conform to a kind of Pareto principle with regard to credence in the truth, and one advantage of the above definition is that it fulfils this principle.

But V is much stronger than the Pareto principle, and essentially corresponds to a kind of utilitarianism with credence taking the role of utility. How can we get to V, or a measure like it, from the Pareto principle?

Generally, a measure of epistemic value can be seen as a function ν defined on individual states of a social network, on practices implementable in such a network, or on parts of these, which takes values among some ordered set (such as the real numbers). As with any kind of value, there are primarily three things we need to know: what is the valuable thing, how much is it worth, and how should we aggregate this thing over different persons, times, states of the world, etc.?

Goldman's suggested answer to what is valuable is that credence in the truth is what we are after, and its worth is equal to its strength. We have agreed to the part of the answer that concerns the "what" question, but this does not by itself settle the "how much" question. This shows up when we combine questions of amount and weighing. For instance, is the state of one person being certain of the truth and one person being certain of falsity really just as good as the state where two persons are completely undecided? Epistemologically, it may be held that the second is far superior, since the people who are undecided still are open to revising their beliefs, while those who are certain of something false are incapable of such revision, at least in the Bayesian model. However, V gives both states the same value.

Centering on the weighing issue, how should we value a network over time? Is it only the 'final' state that matters, or should the credences during all times count? If so, *how* should they count? If time is infinite, there is no way to keep all times equally important while avoiding infinities and incomparabilities. And how should differing populations affect the value? If we take epistemic value to be determined by the *average* credence, we can increase epistemic good by killing off everyone who is not absolutely certain of the truth. This may fairly be held to be a quite unappealing feature of a measure of epistemic value.

These questions are very important but unfortunately also very difficult.

Their difficulty stems primarily because from the complexity of ethical consequentialism, although we in the epistemic case are dealing with a specialization to a certain type of value. However, I think there may be ways forward if we are careful and methodological.

To start with, the most important question to ask is "what do we use epistemic value for?". My proposed answer is that epistemic value is what is maximized by the decisions of an idealized social epistemologist and decision maker (the *evaluator)*, who values knowledge alone. Typical examples of such decisions concern what practices to adopt in a group or society, how such groups are to be assembled, and what rules should be adhered to by inquirers. That the evaluator should value nothing but knowledge entails that any two states of the world (or in our case, the social network) that have the same epistemological properties must be assigned the same value.

By tying epistemic value to epistemic decisions, we open for the possibility of applying decision theory to derive the properties of epistemic value. We do not have the space to go into details of how this may be done here, but a few observations are in order. First of all, decision theory requires that values should be determined up to a positive affine transformation. This, in turn, means that as long as we assume the inquirers involved to be idealized decision makers as well, the *total* value of a practice must be determined fairly strictly from the values of individual states.

For this to work, we first need to determine what we mean by a practice. Let an *individual state* consist of the values of the epistemic variables of a single inquirer α. Let a *network state* be defined as an assignment of values to these variables for all inquirers in a network, as well as for the links, and a *network evolution* as a sequence $E = S^0, S^1, S^2, \ldots$ such that state S^{t+1} is obtainable from the state S^t by having the participants in S^t conditionalize on new information according to the model laid out in the previous sections.

A *practice* π can, to a first approximation, be viewed as a constraint on such evolutions, or, equivalently, as a set of them — those evolutions that are compatible with the practice. But every evolution is determined (at least probabilistically) by its initial state S^0, so we may as well say that the practice is a set of network states, which are to be taken as allowed initial states in applications of the practice.

To be able to evaluate a practice then requires us to weigh over these applications, but decision theory demands that this is to be done by simply adding up the probability-weighted values of the possible evolutions, where the probabilities are those of the evaluator. Thus, assuming P to be her personal probability measure over the space of all possible initial states, the probability to be used is the conditional one $P(E \mid \pi \text{ is adopted})$, which we can expand as

$$P(S^0 \text{ is the actual initial state} \mid \pi \text{ is adopted}) \cdot \prod_{t=0}^{\infty} P(S^{t+1} \mid S^t)$$

for any evolution E. How are we to evaluate evolutions then? Again, decision theory helps. Assuming that not only the evaluator, but also the inquirers are

to be bound by its axioms, we can apply a theorem of Harsanyi [6] to prove that the value of an evolution ϵ must be a *homogenous function* of the values of its individual states. A function $f(x_0, x_1, x_2, \ldots)$ is homogenous if its value, when any of its arguments is multiplied by δ, also is multiplied by δ.

For functions of a finite number of variables, $f(x_1, \ldots, x_n)$ is homogenous iff $f(x_1, \ldots, x_n)$ is a weighted sum of $x_1, \ldots, x_n$. But evolutions are not finite in general, and thus a few more ways of weighing are allowed. One of these is the *limit* function, so the weighing

$$\nu(E) = \lim_{t \to \infty} \nu(S^t)$$

is permissible as well.

There is reason to distinguish between two sorts of practices: the *temporary*, which are implemented so as to have a projected end, and the *continual*, which do not have such an end. An example of the first is an investigation made by a committe, while the second type includes policies such as free speech. For the first, there is for any evolution a specific *final state* S^f, and the difference between this state and the initial state appears reasonable to use as a measure. Thus we assign S^0 the weight -1.0, S^f the weight 1.0, and all other states weight 0. This agrees with Goldman's own suggestion for how to evaluate practices.

For a continual practice, the problem is harder. Although we can use $\lim_{t \to \infty} \nu(S^t)$, it is not certain that this value is very relevant, since we have no guarantee that it is approached except in the infinite limit. It may be that the only reasonable thing to do is to identify some finite period during which the practice is to be evaluated. Doing this introduces another degree of arbitrariness in our evaluation procedure, however.

Disregarding these problems and pressing on, we then come to the evaluation of a single society state S^t in terms of the evaluation of its individual states $\alpha_1^t, \ldots, \alpha_n^t$. Another application of Harsanyi's theorem yields the result that we must use a weighted average here. Although it may seem attractive to assign equal weights to everyone, this is not strictly required, and for some applications it may even be better to assign some of the inquirers zero weight. Such a situation can arise when it is more important for some of the inquirers in the network to have the correct opinion than for others, for example because the former have more power to influence further decisions.

This means that, given a measure ν of epistemic value for an individual at a time, the value of a practice is largely determined by that measure. This, in turn, can be conventionally decided on as an explication of what we mean by *epistemic* value. One way to arrive at Goldman's V is to adopt some axiom such as "adding any equally good amount of certainty in the truth to two states that are equally good preserves their equality" together with structural assumptions. This process essentially gives rise to an extensive measurement structure, which allows certainty in truth to be measured in much the same way as, say, mass or length. The classical reference for this procedure is the trilogy of books on measurement theory by Krantz, Luce, Suppes and Tversky [10].

As advertised, we will not work out the details of how to measure epistemic value here, but I hope that the remarks I have given illustrate that the problem may be approachable in a systematic way. Our present concern is with another aspect of social epistemology: *given* a measure ν of epistemic value, how do we find social networks or practices that maximize the value of this measure? The complexity of this problem itself seems staggering: not only do we have to consider all the possible evolutions of a social network, but for a practice also all the possible initial states it may be applied to. How are we to accomplish this?

4 A Voyage to Laputa

Laputa is a program developed by the author, for use as as a sandbox environment for experiments in social epistemology.[4] It is intended to be useful for investigating several kinds of models of society, but currently it implements the above model, and thus allows us to study the Bayesian social networks presented in this paper.

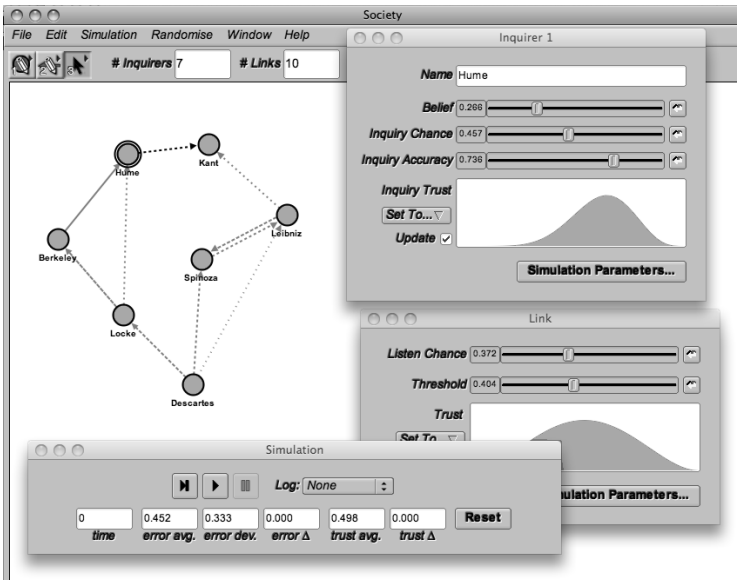


Figure 2: Laputa.

Laputa allows an experimenter to design social networks, to simulate their evolution, and to collect statistics. It also contains methods for evaluating social practices of the kind discussed in the preceding section. Let us, as an example, try to model the editorial process of a scientific journal. Let q be the sentence

[4]The name is taken from the flying city of pseudo-intellectuals in Gulliver's travels. It is freely available for academic purposes, complete with source code, at [. . .].

'The paper currently under review makes enough of a new contribution to be published'. Let p be either q or not-q, depending on whether the paper in question actually meets the criteria or not. We, as evaluators, do not in general know whether $p = q$ or $p = \neg q$, but this does not hinder us from evaluating the performance of the review process.

To create the social network needed for the model we will use three types of inquirers:

- The author. In general, she is quite convinced of q (or she wouldn't have submitted the paper), and thus she is quite convinced either of p or of not-p. We model this by letting her degree of belief in p be selected according to a beta distribution with parameters $\alpha = \beta = 0.5$ (the beta distribution, for these values, has sort of a 'U' shape.) Apart from this, we can not say much about the author, so we let her other properties vary freely.

- The editor. This is an inquirer only in a loose sense of the word, since she in general does not perform any inquiry into the truth of p herself, but relies on the referees for this. We model this by setting the editor's inquiry activity to 0.

- Referees. These are to read the paper, do their own inquiry, and give recommendations and comments to both the editor and the author. In general, referees may be expected to be fairly reliable as inquirers, so while we do not want to guarantee their competence, we model them as having an random inquiry aptitude with a mean of $^2/_3$.

Between these are the links. We can expect the editor to have high trust in the referees (or she wouldn't have consulted them in the first place), but neither the editor not the referees can be taken to know anything about the author's trustworthiness to start with. Putting it together, for two referees, we end up with a social network with the structure depicted in fig. 3.

Now, there are a number of modifications we could make. What if we had more than two reviewers? Are the two we have necessary at all? What if the reviewers were able to send comments to one another, in addition to sending them to the editor and the author? Laputa makes it easy to study these scenarios – all we have to do is to set up the network, and then define an abstract practice to test. The following table lists the influence of three network structures on the editor's credence in the truth as a consequence of the process.[5]

[5]The results were obtained after running 100000 networks for ten steps each. They are statistically significant at the 99% level.

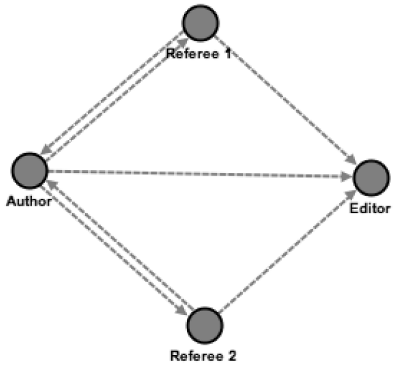


Figure 3: An editorial board.

Number of referees	Referees allowed to communicate	Increase in editor's credence in truth
0	—	0.00
2	Yes	0.10
	No	0.10
4	Yes	0.10
	No	0.12

Looking at the table, we see that the referees definitely help, since without them, the editor has no way to improve her opinion on whether p is the case or not. It is also the case that having 4 reviewers is better than having 2 *when we do not allow them to communicate*. In that case, we thus obtain a sort of justification for the usual practice of not allowing communication among the referees.

This model uses the assumption that while the referees in general are competent inquirers, neither they nor the editor can be expected to have prior knowledge of whether or not p is true, so their prior degrees of belief may vary freely. But there are also cases where the articles sent to a journal tend to be about things that its editor and referees already have informed opinions on. Changing the prior beliefs of the editor and the referees to be randomly spread out according to a beta distribution with a mean at $^2/_3$, we obtain the following result:

Number of referees	Referees allowed to communicate	Increase in editor's credence in truth
0	—	0.08
2	Yes	0.28
	No	0.29
4	Yes	0.31
	No	0.32

As the editor is assumed to have some prior knowledge on p, she can use this to evaluate the author's claim even without relying on referees. But, as the table shows, having referees makes for a much larger increase in her credence in truth. The difference between 2 and 4 referees is not very large, though in both cases not allowing the referees to communicate is somewhat better than doing so.

The recommendation that communication among referees should be disallowed is thus stable over the scenarios we have studied here. But one might hold that we have evaluated the question in the wrong way: we have looked at reduction in the editor's degree of error only. This is reasonable, since after all, it is the editor who makes the decision as to whether a paper should be published or not. But there is also another side to the story. The review process sometimes tends to improve the credences of both the author and the referees as well.

This means that we have reason to take into account *every* participant's credence. Although the question of how to weigh these, as we pointed out in the last section, is far from straightforward, the following table lists the average increase in credence in the truth among all participants in the network, for the scenarios we have studied so far. This value is equal to Goldman's V-value.

Editor and referees have prior knowledge	Number of referees	Referees allowed to communicate	Increase in average credence in truth
Yes	0	—	0.04
	2	Yes	0.28
		No	0.24
	4	Yes	0.33
		No	0.25
No	0	—	0.00
	2	Yes	0.28
		No	0.24
	4	Yes	0.09
		No	0.14

What conclusions can we draw from these results? At least when the re-

viewers and the editor have prior knowledge of p, allowing communication is definitely better than disallowing it, which is contrary to our earlier recommendation. Thus, when we value the quality of articles published only, the currently more common system (which does not allow communication) is preferrable, but when we step back and consider the effects on everyone involved, it may be that we should consider inter-referee communication as well. Furthermore, whether it is better depends on parameters such as the number of reviewers and their prior opinions on p.

Taking another step back, we can ask ourselves not only how the process affects the ones involved, but the readers of the journal as well. This invites us to include more inquirers in the network. While the complexity increases gradually, the work involved is still manageable.

5 Further Adventures in Social Epistemology

The last section gave an example of how to use the model presented here, in conjunction with Laputa, to shed light on problems in social epistemology. This makes the process more craftsmanlike, and perhaps less philosophical. The philosophical work, in turn, is largely displaced to the posing and explication of the problem itself. What model to use is a philosophical as well as mathematical problem. I believe that the one presented here has several useful features, but it also has shortcomings.

For one thing, communication among inquirers is limited to saying p or not-p. There is no such thing as saying 'I have faily good evidence that p', for instance. Inquirers also cannot say 'I believe p because q', which may be held to be an important practice in scientific communication. In Goldman's terms, we have modelled *testimony*, but not *argumentation*, since inquirers are unable to give justification for their claims.

If we were able to model argumentation, this could furthermore allow us to measure the value of stronger senses of knowledge as well. Actually, as the model is now, we can capture parts of the reliabilist conception of knowledge. The number $R_{\sigma\alpha}$, for any source σ, gives a notion of reliability of that source, and we could consider only valuing increases in belief that are produced due to input from sources with high values of $R_{\sigma\alpha}$, or possibly weighing the value of credence by this number. But to some degree, this is already taken into account because our model is statistical. Since a source that has low reliability tends to produce less true belief, it will raise the average degree of belief in truth less than one that has high reliability, and thus typically have less influence on the total epistemic value anyway.

On the other hand, one could hold that on the internalist interpretation, justification has a special value. We do not even need to see it as knowledge, since it appears that having justification for p would increase your *understanding*, and understanding is arguably epistemically valuable whether it is part of knowledge or not (the most well-known contemporary argument for this is made by Kvanvig, cf. [11]). Thus, being able to model justification can allow use of

more detailed measures of epistemic value.

Another limitation is that the model currently only allows discussion of *one* question: whether or not p. If this is a reasonable assumption to make depends on how inquirers were to handle trust in each other with regard to different questions. We can isolate two strategies here: the *generalist* and the *particularist*. A generalist assigns one trust function to each source, and thus treats these as roughly equally trustworthy for every question. A particularist keeps a separate trust function for every question, and can thereby treat any given source as an expert on some questions but not others.

Now, if everyone is a particularist, and the questions are logically independent, no interaction between the inquirers' behavior will occur, and we can model a social network studying n questions by simply modelling n networks, and averaging the results. The interesting cases occur with logical dependence between questions, or in networks with generalist inquirers. For a generalist, a source can build up an inquirer's trust by feeding her expected information (i.e. information that agrees with her prior beliefs), and then use this trust to change her beliefs about something else.

It may be that most people work in the generalist manner, or perhaps according to some mixture of generalism and particularism. Nevertheless, it seems that particularism would be preferrable from an epistemological standpoint, since inquirers generally are not equally reliable when it comes to different areas of expertise. As it also seems like logical dependence in questions can be circumvented by rephrasing them, it may be that the model's limitation to discussion of a single question is quite innocent.

Somewhat less innocent is the assumption of source independence (SI). We have accepted it as an approximation, but it may be that it makes the model deviate too much from actual practice. For one thing, sources are treated both as independent among one another and over time: if β tells α that p twice, α will treat β as having obtained new evidence for p the second time. A side effect of this is that the model will treat nagging as a potentially effective way of spreading one's beliefs. This might be realistic, but it is still a defect.

The way to solve this problem seems to be to model inquirers' beliefs about other inquirers' sources and evidence as well. This, however, adds a large level of complexity, and also requires communication about sources and evidence to be in place. The model, as it is now, seems to be a reasonable approximation: as we mentioned, *we* do not usually keep track of dependences among our sources, so it might be defensible for the inquirers in a social network not to do so either.

Another possibility for refinement concerns the dynamics of a network. The current model handles changes in degrees of belief and in degrees of trust, but there may be occasions where other variables can change over time as well. We have treated the network structure itself as static, but an interesting class of problems in social epistemology concerns how such networks come into being, and studying these requires the links and perhaps even what inquirers are in the network to be subject to change.

More fundamentally, we can question the approrpriateness of Bayesianism and standard decision theory for modelling social epistemology. While there

once seemed to be fairly good evidence that people made their decisions in accordance with its axioms (cf. [2]), later experiments showed that systematic deviations are common as well (cf. [8, 9]). This poses a problem for us: should we model inquirers as they *should* act (in which case Bayesianism seems more easily defensible), or as they actually *do* act (in which it may be better to model belief update in some other way). There is a good case to be made for the second of these alternatives, and this opens up the possibility of a fruitful collaboration between epistemologists, empirical psyhologists and sociologists.

This brief overview should make it clear that I see both the social network model and Laputa itself as proposals for a research program, rather than a finished answer. The aim of such program would be to work out a systematic methodology for concrete problems in social epistemology, such as the one with the review board that we studied. Such a methodology is built on two pillars: modelling and simulation. We approach a social epistemological question by building a formal model of the type of society or group it pertains to. This is done in the first two sections of this paper. To actually draw conclusions, computer simulation is used. It is likely that the complexity of the models required will preclude use of analytical methods, save in certain simple, highly idealized cases. Therefore, a general framework for studying and comparing models is required. It is my hope that Laputa, in time, will evolve into such a framework.

A Derivation of the Credence Update Function

We make the following assumptions, for any inquirer α, any source σ connected to α, any time t, and $\rho \in [0, 1]$.

Principal Principle (PP)

$$C_\alpha^t(S_{\sigma\alpha}p \mid S_{\sigma\alpha} \ \wedge \ R_{\sigma\alpha} = \rho \ \wedge \ p) = \rho$$
$$C_\alpha^t(S_{\sigma\alpha}\neg p \mid S_{\sigma\alpha} \ \wedge \ R_{\sigma\alpha} = \rho \ \wedge \ \neg p) = \rho$$

Communication Independence (CI)

$$C_\alpha^t\left(p \wedge S_{\sigma\alpha}^t \wedge R_{\sigma\alpha} = \rho\right) = C_\alpha^t(p)\ C_\alpha^t(S_{\sigma\alpha}^t)\ \tau_{\sigma\alpha}^t(\rho)$$
$$C_\alpha^t\left(\neg p \wedge S_{\sigma\alpha}^t \wedge R_{\sigma\alpha} = \rho\right) = C_\alpha^t(\neg p)\ C_\alpha^t(S_{\sigma\alpha}^t)\ \overline{\tau}_{\sigma\alpha}^t(\rho)$$

Source Independence (SI)

$$C_\alpha^t\left(\bigwedge_{\sigma\in\Sigma_\alpha^t} S_{\sigma\alpha}^t m_{\sigma\alpha}^t \,\middle|\, p\right) = \prod_{\sigma\in\Sigma_\alpha^t} C_\alpha^t\left(S_{\sigma\alpha}^t m_{\sigma\alpha}^t \mid p\right)$$
$$C_\alpha^t\left(\bigwedge_{\sigma\in\Sigma_\alpha^t} S_{\sigma\alpha}^t m_{\sigma\alpha}^t \,\middle|\, \neg p\right) = \prod_{\sigma\in\Sigma_\alpha^t} C_\alpha^t\left(S_{\sigma\alpha}^t m_{\sigma\alpha}^t \mid \neg p\right)$$

where Σ^t_α is the set of sources that give information to α at t, and $m^t_{\sigma\alpha}$ is what this information says (i.e. p or not-p). The derivation proceeds as follows. By conditionalization, we must have that

$$C^{t+1}_\alpha(p) = C^t_\alpha\left(p \;\middle|\; \bigwedge_{\sigma\in\Sigma^t_\alpha} S_{\sigma\alpha}m^t_{\sigma\alpha}\right)$$

Applying first Bayes' theorem and then (SI) to this expression, we get

$$\begin{aligned}
C^{t+1}_\alpha(p) =& C^t_\alpha\left(p \;\middle|\; \bigwedge_{\sigma\in\Sigma^t_\alpha} S_{\sigma\alpha}m^t_{\sigma\alpha}\right)\\
=& \frac{C^t_\alpha\left(\bigwedge_{\sigma\in\Sigma^t_\alpha} S^t_{\sigma\alpha}m^t_{\sigma\alpha} \;\middle|\; p\right)\, C^t_\alpha(p)}{C^t_\alpha\left(\bigwedge_{\sigma\in\Sigma^t_\alpha} S^t_{\sigma\alpha}m^t_{\sigma\alpha} \;\middle|\; p\right)\, C^t_\alpha(p) + C^t_\alpha\left(\bigwedge_{\sigma\in\Sigma^t_\alpha} S^t_{\sigma\alpha}m^t_{\sigma\alpha} \;\middle|\; \neg p\right)\, C^t_\alpha(\neg p)}\\
=& \frac{\prod_{\sigma\in\Sigma^t_\alpha} C^t_\alpha\left(S^t_{\sigma\alpha}m^t_{\sigma\alpha} \mid p\right)\, C^t_\alpha(p)}{\prod_{\sigma\in\Sigma^t_\alpha} C^t_\alpha\left(S^t_{\sigma\alpha}m^t_{\sigma\alpha} \mid p\right)\, C^t_\alpha(p) + \prod_{\sigma\in\Sigma^t_\alpha} C^t_\alpha\left(S_{\sigma\alpha}m^t_{\sigma\alpha} \mid \neg p\right)\, C^t_\alpha(\neg p)}
\end{aligned}$$

which gives the posterior credence in terms of the values $C^t_\alpha(S^t_{\sigma\alpha}p \mid p)$ and $C^t_\alpha(S^t_{\sigma\alpha}p \mid \neg p)$, for all sources σ. Our next task is thus to derive these expressions. First of all, since $S^t_{\sigma\alpha}p$ is equivalent to $S^t_{\sigma\alpha}p \wedge S^t_{\sigma\alpha}$, $C^t_\alpha(S^t_{\sigma\alpha}p \mid p) = C^t_\alpha(S^t_{\sigma\alpha}p \wedge S^t_{\sigma\alpha} \mid p)$. Applying first the definition of conditional probability and then the theorem of total probability, the definition of conditional probability, and finally (CI), we get

$$\begin{aligned}
C^t_\alpha(S^t_{\sigma\alpha}p \mid p) &= \frac{1}{C^t_\alpha(p)}\, C^t_\alpha(S^t_{\sigma\alpha}p \wedge S^t_{\sigma\alpha} \wedge p)\\
&= \frac{1}{C^t_\alpha(p)} \int_0^1 C^t_\alpha(S^t_{\sigma\alpha}p \wedge S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge p)\, d\rho\\
&= \frac{1}{C^t_\alpha(p)} \int_0^1 C^t_\alpha(S^t_{\sigma\alpha}p \mid S^t_{\sigma\alpha} \wedge R^t_{\sigma\alpha} = \rho \wedge p) C^t_\alpha(S^t_{\sigma\alpha} \wedge R^t_{\sigma\alpha}(p) = \rho \wedge p)\, d\rho\\
&= \frac{1}{C^t_\alpha(p)} \int_0^1 C^t_\alpha(S^t_{\sigma\alpha}p \mid S^t_{\sigma\alpha} \wedge R^t_{\sigma\alpha} = \rho \wedge p) C^t_\alpha(S^t_{\sigma\alpha}) C^t_\alpha(p) \tau^t_\alpha(\rho)\, d\rho\\
&= C^t_\alpha(S^t_{\sigma\alpha}) \int_0^1 C^t_\alpha(S^t_{\sigma\alpha}p \mid S^t_{\sigma\alpha} \wedge R^t_{\sigma\alpha}(p) = \rho \wedge p) \tau^t_\alpha(\rho)\, d\rho
\end{aligned}$$

But because of (PP), $C^t_\alpha(S^t_{\sigma\alpha}p \mid S^t_{\sigma\alpha} \wedge R^t_{\sigma\alpha}(p) = \rho \wedge p) = \rho$, so we get

$$C^t_\alpha(S^t_{\sigma\alpha}p \mid p) = C^t_\alpha(S^t_{\sigma\alpha}) \int_0^1 \rho\ \tau^t_\alpha(\rho)\ d\rho = C^t_\alpha(S^t_{\sigma\alpha})\ \langle\tau^t_{\sigma\alpha}\rangle$$

Parallel derivations give that

$$\begin{aligned} C^t_\alpha(S^t_{\sigma\alpha}\neg p \mid p) &= C^t_\alpha(S^t_{\sigma\alpha})\ \langle\overline{\tau}^t_{\sigma\alpha}\rangle \\ C^t_\alpha(S^t_{\sigma\alpha}p \mid \neg p) &= C^t_\alpha(S^t_{\sigma\alpha})\ \langle\tau^t_{\sigma\alpha}\rangle \\ C^t_\alpha(S^t_{\sigma\alpha}\neg p \mid \neg p) &= C^t_\alpha(S^t_{\sigma\alpha})\ \langle\overline{\tau}^t_{\sigma\alpha}\rangle \end{aligned}$$

Plugging this into our earlier result gives

$$\begin{aligned} C^{t+1}_\alpha(p) &= \frac{C^t_\alpha(p) \prod\limits_{\sigma\in\Sigma^t_\alpha} C^t_\alpha(S^t_{\sigma\alpha})\ \langle\tau^t_{\sigma\alpha}\rangle}{C^t_\alpha(p) \prod\limits_{\sigma\in\Sigma^t_\alpha} C^t_\alpha(S^t_{\sigma\alpha})\ \langle\tau^t_{\sigma\alpha}\rangle + C^t_\alpha(\neg p) \prod\limits_{\sigma\in\Sigma^t_\alpha} C^t_\alpha(S^t_{\sigma\alpha})\ \langle\overline{\tau}^t_{\sigma\alpha}\rangle} \\ &= \frac{C^t_\alpha(p) \prod\limits_{\sigma\in\Sigma^t_\alpha} \langle\tau^t_{\sigma\alpha}\rangle}{C^t_\alpha(p) \prod\limits_{\sigma\in\Sigma^t_\alpha} \langle\tau^t_{\sigma\alpha}\rangle + C^t_\alpha(\neg p) \prod\limits_{\sigma\in\Sigma^t_\alpha} \langle\overline{\tau}^t_{\sigma\alpha}\rangle} \end{aligned}$$

which is the result sought.

B Derivation of the Trust Update Function

Again, we assume (PP) and (CI) of the last appendix to hold (although we do not need (SI) for this derivation). The value we wish to derive is

$$\tau^{t+1}_{\sigma\alpha} = \frac{d}{d\rho}\ C^{t+1}_\alpha(R_{\sigma\alpha} = \rho) = \frac{d}{d\rho}\ C^t_\alpha(R_{\sigma\alpha} = \rho \mid S^t_{\sigma\alpha}m^t_{\sigma\alpha})$$

for an arbitrary source σ of α, and an arbitrary message $m^t_{\sigma\alpha}$ given by σ to α at t. For simplicity, assume that $m^t_{\sigma\alpha} = p$ (the case where $m^t_{\sigma\alpha} = \neg p$ is completely parallel). Applying the definition of conditional probability, the equivalence $S^t_{\sigma\alpha} \wedge S^t_{\sigma\alpha}p \Leftrightarrow S^t_{\sigma\alpha}p$, and then the theorem of total probability, we get

$$\begin{aligned} C^t_\alpha(R_{\sigma\alpha} = \rho \mid S^t_{\sigma\alpha}p) &= \frac{C^t_\alpha(R_{\sigma\alpha} = \rho \wedge S_{\sigma\alpha}p)}{C^t_\alpha(S_{\sigma\alpha}p)} = \frac{C^t_\alpha(R_{\sigma\alpha} = \rho \wedge S^t_{\sigma\alpha}p \wedge S^t_{\sigma\alpha})}{C^t_\alpha(S_{\sigma\alpha}p)} \\ &= \frac{1}{C^t_\alpha(S^t_{\sigma\alpha}p)}\ C^t_\alpha(R_{\sigma\alpha} = \rho \wedge S^t_{\sigma\alpha}p \wedge S^t_{\sigma\alpha} \wedge p) + C^t_\alpha(R_{\sigma\alpha} = \rho \wedge S^t_{\sigma\alpha}p \wedge S^t_{\sigma\alpha} \wedge \neg p) \\ &= \frac{1}{C^t_\alpha(S^t_{\sigma\alpha}p)} \Big(C^t_\alpha(S^t_{\sigma\alpha}p \mid R_{\sigma\alpha}(p) \wedge S^t_{\sigma\alpha} \wedge p)\ C^t_\alpha(R_{\sigma\alpha} = \rho \wedge S^t_{\sigma\alpha} \wedge p) \\ &\qquad + C^t_\alpha(S^t_{\sigma\alpha} \mid R_{\sigma\alpha} = \rho \wedge S^t_{\sigma\alpha} \wedge \neg p)\ C^t_\alpha(R_{\sigma\alpha} = \rho \wedge S^t_{\sigma\alpha} \wedge \neg p)\Big) \end{aligned}$$

Now, we apply (PP) and (CI), and then again the equivalence $S^t_{\sigma\alpha} \wedge S^t_{\sigma\alpha} p \Leftrightarrow S^t_{\sigma\alpha} p$:

$$
\begin{aligned}
&C^t_\alpha(R_{\sigma\alpha} = \rho \mid S^t_{\sigma\alpha} p) \\
&\qquad = C^t_\alpha(R_{\sigma\alpha} = \rho)\; C^t_\alpha(S^t_{\sigma\alpha})\; \frac{\rho\; C^t_\alpha(p) + (1-\rho)\; C^t_\alpha(\neg p)}{C^t_\alpha(S_{\sigma\alpha} p)} \\
&\qquad\qquad = C^t_\alpha(R_{\sigma\alpha} = \rho) \frac{\rho\; C^t_\alpha(p) + (1-\rho)\; C^t_\alpha(\neg p)}{C^t_\alpha(S^t_{\sigma\alpha} p \mid S^t_{\sigma\alpha})}
\end{aligned}
$$

How do we calculate the denominator $C^t_\alpha(S^t_{\sigma\alpha} p \mid S^t_{\sigma\alpha})$ here? We use the definition of conditional probability and expand twice using the theorem of total probability:

$$
\begin{aligned}
C^t_\alpha(S^t_{\sigma\alpha} p \mid S^t_{\sigma\alpha}) &= \frac{C^t_\alpha(S^t_{\sigma\alpha} p \wedge S^t_{\sigma\alpha})}{C^t_\alpha(S^t_{\sigma\alpha})} \\
&= \frac{C^t_\alpha(S^t_{\sigma\alpha} p \wedge S^t_{\sigma\alpha} \wedge p) + C^t_\alpha(S^t_{\sigma\alpha} p \wedge S^t_{\sigma\alpha} \wedge \neg p)}{C^t_\alpha(S^t_{\sigma\alpha})} \\
= \frac{1}{C^t_\alpha(S^t_{\sigma\alpha})} \int_0^1 & C^t_\alpha(S^t_{\sigma\alpha} p \wedge S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge p) + C^t_\alpha(S^t_{\sigma\alpha} p \wedge S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge \neg p)\; d\rho \\
&= \frac{1}{C^t_\alpha(S^t_{\sigma\alpha})} \int_0^1 C^t_\alpha(S^t_{\sigma\alpha} p \mid S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge p)\; C^t_\alpha(S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge p) \\
&\qquad + C^t_\alpha(S^t_{\sigma\alpha} p \mid S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge \neg p)\; C^t_\alpha(S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge \neg p)\; d\rho
\end{aligned}
$$

Applying (PP) and (CI) to this expression then yields

$$
\begin{aligned}
C^t_\alpha(S^t_{\sigma\alpha} p \mid S^t_{\sigma\alpha}) = \frac{1}{C^t_\alpha(S^t_{\sigma\alpha})} \Big(& C^t_\alpha(p)\; C^t_\alpha(S^t_{\sigma\alpha}) \int_0^1 \rho\; C^t_\alpha(R_{\sigma\alpha} = \rho)\; d\rho + \\
& + C^t_\alpha(\neg p)\; C^t_\alpha(S^t_{\sigma\alpha}) \int_0^1 (1-\rho)\; C^t_\alpha(R_{\sigma\alpha} = \rho)\; d\rho \Big) \\
&= C^t_\alpha(p)\; \langle \tau^t_{\sigma\alpha} \rangle + C^t_\alpha(\neg p) \langle \overline{\tau}^t_{\sigma\alpha} \rangle
\end{aligned}
$$

Putting it all together, we finally get

$$
\tau^{t+1}_{\sigma\alpha} = \frac{d}{d\rho}\; C^t_\alpha(R_{\sigma\alpha}(p) \mid S^t_{\sigma\alpha} p) = \tau^t_{\sigma\alpha}(\rho) \frac{\rho\; C^t_\alpha(p) + (1-\rho)\; C^t_\alpha(\neg p)}{C^t_\alpha(p)\; \langle \tau^t_{\sigma\alpha} \rangle + C^t_\alpha(\neg p) \langle \overline{\tau}^t_{\sigma\alpha} \rangle}
$$

References

[1] Baltag, Alexandru, Lawrence S. Moss and Slawomir Solecki. 1998. "The Logic of Public Announcements, Common Knowledge and Private Suspicions". *Proceedings of TARK'98* (Seventh Conference on Theoretical Aspects of Rationality and Knowledge).

[2] Davidson, Donald, Patrick Suppes and Sidney Siegel. 1959. *Decision Making: An Experimental Approach*. Stanford University Press.

[3] van Ditmarsch, Hans, Wiebe van der Hoek and Barteld Kooi. 1998. *Dymanic Epistemic Logic*. Springer.

[4] Goldman, Alvin I. 1986. *Epistemology and Cognition*. Harvard University Press.

[5] — 1999. *Knowledge in a Social World*. Oxford University Press, 1999.

[6] Harsanyi, John. 1955. "Cardinal Welfare, Individualistic Ethics and Interpersonal Comparisons of Utility'. *Journal of Political Economy* 63: 309–21.

[7] Jeffrey, Richard C. 1983 *The Logic of Decision*, 2nd ed. University of Chicago Press.

[8] Kahneman, Daniel and Amos Tversky, "Prospect Theory: an Analysis of Decision under Risk". *Econometrica* 47: 263–91.

[9] Kahneman Daniel, Paul Slovic and Amos Tversky (eds.). 1982. *Judgment Under Uncertainty. Heuristics and Biases*. Cambridge University Press.

[10] Krantz, David H., R. Duncan Luce, Patrick Suppes and Amos Tversky. 1971, *Foundations of Measurement*, vol. I. Academic Press.

[11] Kvanvig, Jomathan. 2003. *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press.

[12] Lewis, David. 1980. "A Subjectivist's Guide to Objective Chance", in Richard C. Jeffrey (ed.) *Studies in Inductive Logic and Probability*, vol. 2. University of California Press.