A Bayesian Simulation Model of Group Deliberation and Polarization

Erik J. Olsson

Abstract: The paper describes a simulation environment for epistemic interaction based on a Bayesian model called Laputa. An interpretation of the model is proposed under which the exchanges taking place between inquirers are argumentative. The model, under this interpretation, is seen to survive the polarization test: if initially disposed to judge along the same lines inquirers in Laputa will adopt a more extreme position in the same direction as the effect of group deliberation, just like members of real argumentative bodies. Our model allows us to study what happens to mutual trust in the polarization process. We observe that inquirers become increasingly trusting which creates a snowball effect. We also study conditions under which inquirers will diverge and adopt contrary positions. To the extent that Bayesian reasoning is normatively correct, the bottom line is that polarization and divergence are not necessarily the result of mere irrational "group think" but that even ideally rational inquirers will predictably polarize or diverge under realistic conditions. The concluding section comments on the relation between the present model and the influential and empirically robust Persuasive Argument Theory (PAT), and it is argued that the former is essentially subsumable under the latter.

1. Introduction

There has been a lot of experimental work in social psychology of group deliberation and some striking results as well (for an overview of early work, see Isenberg 1986). However, there does not seem to be much work focusing on computer simulation of deliberative processes taking the role of argumentation seriously.[1] This is unlike many other areas studying complex systems, including economics, where simulation models abound. It is easy to understand why they are found to be so useful: the speed at which a computer simulation can be carried out should be compared with the sometimes many months required for meticulously planning and executing a controlled experiment; and, moreover, computer simulations allow for precise control for parameters that can be extremely difficult to control for in real experiments. This increased speed and control is gained, obviously, at the expense of realism because simulation models

[1] For an overview of exact models of opinion dynamics see Hegselmann and Krause (2006). See also Zollman (2007).

need to be idealized in order to be computationally workable. Laboratory experimentation and computer simulation are therefore complementary activities.

This paper contributes to the study of simulation models of group deliberation with the aim of expanding the methodological toolkit available to researchers studying argumentation in a social setting. The model, called Laputa, allows for studying not only the dynamics of belief but also of trust, including mutual trust among inquirers. Laputa was developed by Staffan Angere and the author, with Angere being the main originator and also the programmer behind the simulation environment with the same name. The plan of the paper is as follows. In section 2, I describe Laputa as a simulation framework of epistemic interaction, postponing the description of the underlying Bayesian model until section 3. In section 4, an interpretation is proposed according to which inquirers in Laputa exchange novel arguments on a common issue. I proceed, in section 5, to test whether this model of deliberation exhibits polarization effects. Conditions under which inquirers diverge are also studied. In the concluding section, I comment on the relation between the present model and the influential Persuasive Argument Theory (PAT).

2. The Laputa simulation framework

I will choose to introduce Laputa as a simulation framework leaving the details of the underlying model for the next section. Social networks are represented and depicted in the program as directed graphs in which the nodes represent inquirers and the links represent communication channels (figure 1).

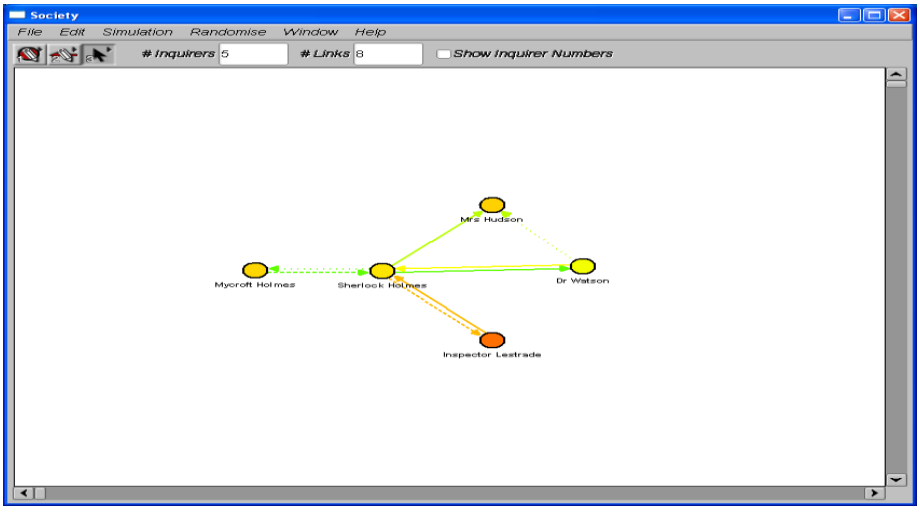

Figure 1: The social network of Sherlock Holmes as represented in Laputa.

A number of parameters can be set for each inquirer. The *initial degree of belief* is the inquirer's initial credence in proposition *p*. *Inquiry accuracy* is the reliability of the inquirer's own inquiries. The *inquiry chance* is the probability that the inquirer will conduct an inquiry. The *inquiry trust* is the inquirer's degree of trust in her own inquiries. Likewise, there are a number of parameters for each link. The *listen trust* is the recipients trust in the sender. The *threshold of assertion* is the degree of confidence in a proposition ("*p*" or "not-*p*") required for the sender to submit a corresponding message to the recipient(s). Whether a message will then be submitted depends on the *listen chance*. For instances, if the threshold is set at .90, this means that the sender needs to believe *p* (not-*p*) to a degree .90 in order for her to send a positive (negative) message in the network.

Running Laputa can mean to construct a network, such as that in figure 1, assign initial values to the inquirer and link parameters, and then click on a "run" button. What happens then is that Laputa runs through a series of steps, each step representing a chance for an inquirer to conduct an inquiry, to communicate (send, listen) to the other inquirers to which she is "hooked up", or to do both. After each step, Laputa will update the whole network according to the information received by the inquirers in accordance with the Bayesian model with which we shall soon become acquainted. Thus, a new degree of belief is computed for each inquirer based on the old degree of belief and the new information received through inquiry and/or listening to other inquirers. Laputa also updates the inquiry trust and listen trust parameters in accordance with Bayesian principles.

Laputa outputs not just what happens to the individual inquirers during simulation, but also collects some statistical data. Thus, *error delta* is the difference between the initial and final average degrees of belief in the proposition *p*, which is assumed true by convention. Given error delta, we can compute the veritistic value (V-value) in the sense of Goldman (1999) for a network evolution according to the following simple rule: V-value = -error delta. This means that an error delta of -0.076 equals a V-value of 0.076. Angere (forthcoming), Olsson (2011) and Olsson and Vallinder (in press) discuss various applications of Laputa relating to Goldman's veritistic social epistemology. See Vallinder and Olsson (in press b) for a further philosophical application of Laputa.

Laputa also allows its user to specify various features or "desiderata" of networks at an abstract level. The program can then randomly generate a large number of networks of different sizes having those features, letting them evolve while collecting various statistics. This is done in Laputa's "batch window" (figure 2), the perhaps most powerful feature of the program.

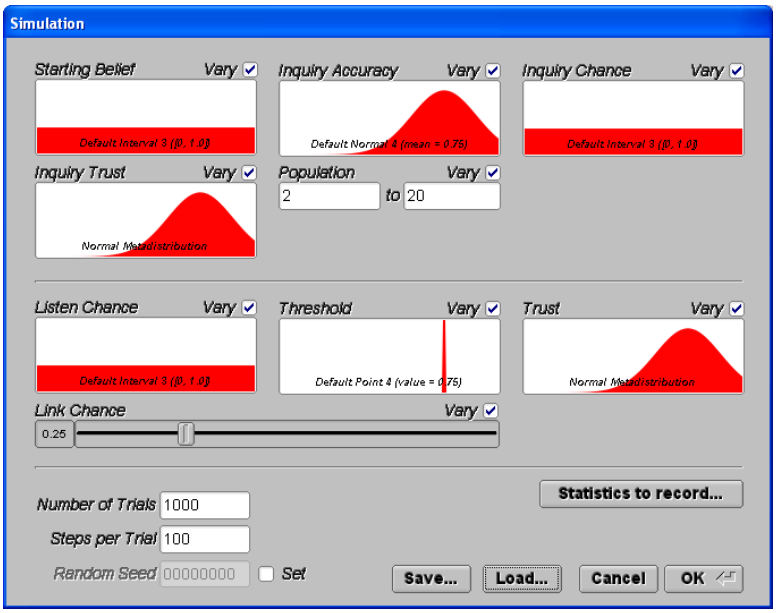

Figure 2: The batch window in Laputa.

In the batch window, various probability distributions can be selected for the several inquirer and link parameters. For instance, the flat distribution for "starting belief" indicates that Laputa, when selecting the initial credences in *p* for a generated network, will treat all possible credences as being equally likely to be realized. The selection of a normal distribution for "Inquiry accuracy", centered around 0.75 means that Laputa, when selecting the inquiry accuracy for the inquirers in the generated networks, will have a preference for assigning an accuracy of 0.75 and surrounding values. The population feature allows the specification of the lower and upper sizes of the networks to be examined. In this case, Laputa is instructed to generate and study networks having 2 to 20 inquirers. "Link chance" specifies the "density" of the networks to be studied. A link chance of 0.25 indicates a 25 percent chance that two inquirers will be connected by a directed communication link. In Figure 2, the number of trials has been set to 1,000, meaning that Laputa will generate and study 1,000 networks in accordance with the statistical criteria specified in the batch window. Finally, the number of steps per trial has been set to 100, indicating that the focus is inquirer interaction over a longer period of time.

3. The underlying Bayesian model

It is time to elucidate the model underlying the simulation environment. This section follows the exposition in Angere (forthcoming), except in one main respect: unlike Angere, I will describe the model in a way that does not presuppose any more specific interpretation of the exchanges

taking place between inquirers. Formally, we can take a *social network S* to be a set Γ of *inquirers*, together with a binary relation *R* on Γ, which we call the *network structure*. This means that, abstractly speaking, a social network is a directed graph.

Following Bayesian tradition, the epistemic state of a person α at time *t* is assumed to be given by a *credence function* $C_\alpha^t : L \rightarrow [0,1]$. L can be taken to be a classical propositional language, and C_α^t is assumed to fulfill the standard axioms of a probability measure. 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. 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*.

In our model, there are two fundamentally different ways for the inquirers to receive new information: 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 others in the network. Paradigmatic cases of inquiry include observation, experiment and taking advice from persons outside the social network.

Not all participants' approaches to inquiry are the same, and they tend to vary both in their degree of activity and their effectiveness. We say that a result of inquiry is *positive* if it supports *p*, and *negative* if it supports not-*p*. Let $S_{i\alpha}^{t+}$ be the proposition "*α*'s inquiry gives a positive result at time *t*", $S_{i\alpha}^{t-}$ be the proposition "*α*'s inquiry gives a negative at *t*", and $S_{i\alpha}^{t} = S_{i\alpha}^{t+} \vee S_{i\alpha}^{t-}$ the proposition that *α*'s inquiry gives *some* result at *t*, positive or negative. We represent the participants' properties *qua* inquirers by two probabilities: the chance $P(S_{i\alpha}^{t})$ that, at any moment *t*, *α* receives a result from her inquiries, and the chance $P(S_{i\alpha}^{t+} | S_{i\alpha}^{t} \wedge p)$ that, when such a result is obtained, it is the right one. To simplify matters, we assume that the chance that inquiry gives an appropriate result does not depend on whether *p* is true or false.

$P(S_{i\alpha}^{t})$ will be referred to as *α*'s *activity*, and $P(S_{i\alpha}^{t+} | S_{i\alpha}^{t} \wedge p)$ as her *aptitude*. An inquirer without interest in *p* would generally have a low activity value, while one very interested in *p*, but engaging in inquiry using faulty methods would have a high activity value 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. Analogously to the inquiry notation we define

$S_{\beta\alpha}^{t+} =_{df} \beta$ sends a positive message to *α* at *t*

$S_{\beta\alpha}^{t-} =_{df} \beta$ sends a negative message to *α* at *t*

$S_{\beta\alpha}^{t} =_{df} \beta$ sends a positive or a negative message to *α* at *t*

This strength of a link $\beta\alpha$ is then representable as a probability $P(S_{\beta\alpha})$ being the chance that β sends some message, whether positive or negative, to α.

Given that β communicates with α, what does she say? And what makes her say it? We will leave the first question for the next section. The second question can be answered by referring to a property of the link $\beta\alpha$ that we will call its *threshold of assertion*: a value $T_{\beta\alpha}$ between 0 and 1, such that

If $T_{\beta\alpha} > 0.5$, β sends a positive message to α only if $C_\beta(p) \geq T_{\beta\alpha}$, and a negative message only if $C_\beta(p) \leq 1 - T_{\beta\alpha}$;
If $T_{\beta\alpha} < 0.5$, β sends a positive message to α only if $C_\beta(p) \leq T_{\beta\alpha}$, and a negative message only if $C_\beta(p) \geq 1 - T_{\beta\alpha}$; and
If $T_{\beta\alpha} = 0.5$, β sends a positive or a negative message to α independently of what she believes, which is modeled by letting her pick what to say randomly.

So far we have described how the inquirers in a social network engage in inquiry and communication, but we have said nothing about how they react to the results of these practices. The purpose of the following considerations is to provide enlightenment in this regard.

We define the *reliability* of α's source σ as

$$R_{\sigma\alpha} =_{df} P(S^+_{\sigma\alpha} | S_{\sigma\alpha} \wedge p) = P(S^-_{\sigma\alpha} | S_{\sigma\alpha} \wedge \neg p)$$

This definition presupposes that the probability that any source sends a positive message, if p is the case, is equal to the probability that it sends a negative message, 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. For a discussion, see Olsson (2011).

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] → [0,1], called *α's trust function for σ at t*, such that

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

for a, b in [0,1]. $\tau_{\sigma\alpha}(\rho)$ 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 $1 - \tau^t_{\sigma\alpha}$, representing α's credence density for propositions about σ *not* being reliable, which we will refer to as $\bar{\tau}^t_{\sigma\alpha}$.

It is reasonable to think 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 as the *principal principle* (Lewis 1980). It 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:

$$C^t_\alpha(S^{t+}_{\sigma\alpha} | S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge p) = \rho$$

$$C^t_\alpha(S^{t-}_{\sigma\alpha} | S^t_{\sigma\alpha} \wedge R_{\sigma\alpha} = \rho \wedge \neg p) = \rho$$

for all t, i.e. α's credence in σ giving a positive report, 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 modeling considerably. The independence assumption we use here will be referred to as *communication independence* (CI):

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

Communication independence implies that whether σ says anything is independent of whether p is actually true as well as of σ´s reliability.

Given (PP) and (CI) we can now define the following expression for α's credence in σ's reliability (see Angere, forthcoming, for the derivation):

$$C^t_\alpha(S^{t+}_{\sigma\alpha} | p) = C^t_\alpha(S^t_{\sigma\alpha}) \int_0^1 \rho \tau^t_{\sigma\alpha}(\rho) d\rho$$

The integral in this expression is the expected value $\langle \tau^t_{\sigma\alpha} \rangle$ of the trust function $\tau^t_{\sigma\alpha}$, whence

$$(*)\ C^t_\alpha(S^{t+}_{\sigma\alpha} | p) = C^t_\alpha(S^t_{\sigma\alpha}) \langle \tau^t_{\sigma\alpha} \rangle$$

Similarly,

(**) $C_\alpha^t(S_{\sigma\alpha}^{t+}|\neg p) = C_\alpha^t(S_{\sigma\alpha}^t)\langle\bar{\tau}_{\sigma\alpha}^t\rangle$

We can now derive the crucial expressions $C_\alpha^t(p|S_{\sigma\alpha}^{t+})$ and $C_\alpha^t(p|S_{\sigma\alpha}^{t-})$, the credence an inquirer should place in p at t given that she receives a positive or a negative message, respectively, from a single source σ:

$$C_\alpha^t(p|S_{\sigma\alpha}^{t+}) = \frac{C_\alpha^t(p)\langle\tau_{\sigma\alpha}^t\rangle}{C_\alpha^t(p)\langle\tau_{\sigma\alpha}^t\rangle + C_\alpha^t(\neg p)\langle\bar{\tau}_{\sigma\alpha}^t\rangle}$$

$$C_\alpha^t(p|S_{\sigma\alpha}^{t-}) = \frac{C_\alpha^t(p)\langle\bar{\tau}_{\sigma\alpha}^t\rangle}{C_\alpha^t(p)\langle\bar{\tau}_{\sigma\alpha}^t\rangle + C_\alpha^t(\neg p)\langle\tau_{\sigma\alpha}^t\rangle}$$

where $\langle\tau_{\sigma\alpha}^t\rangle$ is the expected value of the trust function $\tau_{\sigma\alpha}^t$. By the Bayesian requirement of conditionalization, we must have $C_\alpha^{t+1} = C_\alpha^t(p|S_{\sigma\alpha}^{t+})$, whenever σ is the only source giving information to α at t. This means that these formulae completely determine how α should update her credence in such a case.

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 but we will not go into the details here. A full account can be found in Angere (forthcoming). Suffice it to mention the following consequence of our model: Even if an inquirer happens to be a perfect inquirer insofar as her inquiry always gives the right result, a fairly low stability of her faith in inquiry, together with her prior judgment that p is unlikely, may conspire to make her distrust her own inquiry. This, in turn, may give 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.

The present model gives rise to a number of qualitative updating rules in the case of one message received. We say an inquirer *trusts* a given source if the inquirer's credence in the reliability of the source is greater than 0.5; *distrusts* the source if it is less than 0.5; and *neither trusts nor distrusts* the source otherwise. We say that a message that p (not-p) was surprising to an inquirer if, prior to receiving the message, the inquirer's credence in p (not-p) was less than 0.5; *expected* if it was greater than 0.5; and *neither surprising nor expected* otherwise. In table 1, a +-sign in the left component of a pair (_, _) means that the inquirer's current belief is reinforced (i.e. her credence in the conclusion is strengthened if above 0.5 and weakened if below 0.5). A –-sign means that the inquirer's current belief is weakened (i.e. her credence in the conclusion is weakened if above 0.5 and strengthened if below 0.5), whereas 0 means that the inquirer's credence in the conclusion is left unchanged. A +-sign in the right component of a pair (_, _)

signifies that the juror's trust in the source (i.e. credence in its reliability) is strengthened, a –- sign that it is weakened, and 0 that it is left unchanged. Table 1 shows how updating on the information from one source affects an inquirer under various circumstances.

	Message expected	**Neither nor**	**Message surprising**
Source trusted	(+, +) (a)	(+, 0) (b)	(-, -) (c)
Neither nor	(0, +) (d)	(0, 0) (e)	(0, -) (f)
Source distrusted	(-, +) (g)	(-, 0) (h)	(+, -) (i)

Table 1: Single message updating in Laputa (for credences strictly between 0 and 1).

Suppose, for example, that inquirer α's prior credence in p is 0.7. Now α receives a positive message, i.e. a message in support of p, from β, who we assume to be trusted by α. Since the message is expected and the source is trusted, we have the situation described in cell (a) in table 1. Accordingly, α will react by raising both her degree of belief in p and her degree of trust in β. If, by contrast, the message sent by β is negative, we have the situation depicted in cell (c), so that α will respond by lowering both her degree of belief in p and her trust in β.[2]

Let Δ_α^t be the set of all sources from which α receives information at t. Our Bayesian framework requires that credences be updated by means of conditionalization:

(Cond) $C_\alpha^{t+1}(p) = C_\alpha^t(p \mid \wedge S_{\sigma\alpha}^{t*})$,

where the conjunction runs over all σ in Δ_α^t. $S_{\sigma\alpha}^{t*}$ is the message that α receives from σ at t, i.e., either $S_{\sigma\alpha}^{t+}$ or $S_{\sigma\alpha}^{t-}$. The right hand side of (Cond) can be very hard to assess in the absence of further assumptions. We can simplify the situation considerably by assuming *source independence* (SI):

$C_\alpha^t(\wedge S_{\sigma\alpha}^{t*} \mid p) = \prod C_\alpha^t(S_{\sigma\alpha}^{t*} \mid p)$

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

[2] For proofs of the results summarized in Table 1 see Vallinder and Olsson (in press a)), which is a detailed study of the dynamics of trust in the Laputa model.

Source independence states that the information coming from the sources is independent conditional on the truth as well as on the falsity of *p*. This is the standard Bayesian way of capturing the idea that there is no direct influence between the sources, e.g., that they have not conspired to give a certain message (see, for instance, chapter 2 in Olsson 2005). Given source independence, we can relatively easily compute the left-hand side of (Cond) by relying on Bayes' theorem together with the theorem of total probability. See the appendix to this paper for an example of how the machinery works and of the important role played by the assumption of source independence in the updating of credences.

The bottom line is that, given the complexity of the subject matter, Laputa is a simple and workable model once we assume source independence. But that, of course, is a technical motivation and not a philosophical one. The question is whether an interpretation of Laputa can be found under which source independence is true or at least highly plausible. This is the issue to which we now turn.

4. Interpreting Laputa

In order to be informative, an interpretation of Laputa should say something more precise about what kind of messages inquirers receive from inquiry and from the other inquirers. On what I will call the *opinion disclosure* interpretation of the model the positive messages are simply messages to the effect that *p* is the case, and the negative messages that not-*p* is the case. We let $S_{i\alpha}^{t+}$ be the proposition "α's inquiry signaled that *p* is the case at time *t*", $S_{i\alpha}^{t-}$ be the proposition "α's inquiry signaled that not-*p* is the case at *t*", and $S_{i\alpha}^{t} = S_{i\alpha}^{t+} \vee S_{i\alpha}^{t-}$ the proposition that α's inquiry signaled either that *p* or that not-*p* is the case at *t*. Similarly,

$S_{\beta\alpha}^{t+} =_{df} \beta$ disclosed her opinion that *p* to α at *t*

$S_{\beta\alpha}^{t-} =_{df} \beta$ disclosed her opinion that not-*p* to α at *t*

$S_{\beta\alpha}^{t} =_{df} \beta$ disclosed her opinion that *p* or that not-*p* to α at *t*

Thus what happens, at a given point in a social network evolution, is that one or more inquirers receive messages to the effect that *p* (not-*p*) is the case from their own inquiries and/or from the other inquirers. Social network interaction on this interpretation consists largely in repeated disclosure of opinions. The opinions are disclosed only to those other inquirers with whom the inquirer can communicate. The inquirers then update their credence in *p* at each round by conditionalization in the manner described above.

This was the original interpretation of Laputa as laid out in Angere (forthcoming) and Olsson (2011). Under it, Laputa can be used, at least in principle, for studying the *mere exposure effect* in

social psychology, the claim being that mere exposure to other group members' positions on some issue can move a given member's credence in similar directions (for an overview, see Isenberg 1986, pp. 1142-1144). However, there is a problem with this interpretation which needs to be mentioned. Suppose an inquirer is repeatedly exposing other inquirers to her opinion without her receiving any new information from inquiry in the meantime. Suppose, for example, that she repeatedly informs the others that her opinion is that p is true in consecutive steps of the deliberation. As Laputa is built, this will typically lead the other inquirers to repeatedly update their credence in p in a positive direction and to adopt an ever increasing trust in the discloser. While this effect may be of little statistical significance in the end, it is certainly counterintuitive.

There is another interpretation which does not have this problem. On this interpretation what are exchanged among the inquirers are not opinions but *arguments*. More precisely, inquirers exchange arguments for or against the proposition *p*. Since Laputa does not represent the structure of arguments, this interpretation is in some need of justification.

Our starting point will be the assumption that deliberation, as studied here, is *cooperative*, much in the sense of Grice's maxims for cooperative communication (Grice 1975). Thus, we assume that inquirers adhere to the Maxims of Quality and of Relation. The former states that one should not convey what is believed to be false or unjustified. According to the latter, one should make contributions that are relevant. Giving an invalid argument or an argument with false premises would be in violation of the Maxim of Quality. Cooperative communication requires that all arguments be sound, i.e. valid and based on true premises at least in the eyes of the proponent. In this paper, we will take the inquirers' competence in this regard for granted.

The internal structure of arguments is important if the arguments presented can fail to be sound. The receiver can then determine whether the argument is valid and based on true premises by identifying the argument structure, including the premises and the mode of inference (deductive, inductive etc.). But if all arguments presented are sound, as we have assumed them to be, then it is less obvious that argument structure is of statistical importance. The assumption of soundness can therefore be used to motivate viewing arguments as "black boxes" without any internal structure. What is important in an argument, from this perspective, is whether it is a pro or a con argument vis-à-vis the issue at stake. From this perspective, the Laputa model makes sense as a simplified and idealized model of argumentation.

We have yet to explain why the problem of repetition does not arise under this interpretation. The key idea is to think of the arguments that are put forward by inquirers in Laputa as *novel* arguments, i.e. arguments that have not been advanced earlier in the deliberation process. Hence, if an inquirer repeatedly argues that p this should not be interpreted as the inquirer repeating the same argument for p, but as her advancing a series of

novel arguments to that conclusion. If so, the fact that the inquirers on the receiving side will repeatedly update their credence in the conclusion and their trust in the proponent is not unreasonable. On the contrary, it is what one would expect should happen.

The assumption of novelty can be justified as follows. Kaplan (1977) found that, if arguments are presented that the individual group member is already aware of, a shift in his or her position will not occur as a result of the discussion. The stating of the argument will be seen as an irrelevant deliberative contribution. Vinograd and Bernstein (1978) report similar findings. In other words, giving an argument which has already been taken into account violates the Maxim of Relation, which we have assumed that the inquirers adhere to. If the network is fully connected so that every argument is presented to everyone, there will be common knowledge about which arguments have already been presented. Hence, only novel arguments will be advanced. (If the network is not fully connected, we adopt the same assumption – that all arguments presented are novel – as a useful idealization.)

We say that an argument is *positive* if its conclusion is p, and that it is *negative* if its conclusion is not-p. Putting together what was said above, the proposal is that we take $S_{i\alpha}^{t+}$ to mean "α's inquiry produced a novel positive argument at time t", $S_{i\alpha}^{t-}$ to mean "α's inquiry produced a novel negative argument at t", and $S_{i\alpha}^{t} = S_{i\alpha}^{t+} \vee S_{i\alpha}^{t-}$ to mean that α's inquiry produced some novel argument, whether positive or negative, at t. Similarly,

$S_{\beta\alpha}^{t+} =_{df} \beta$ presented a novel positive argument to α at t

$S_{\beta\alpha}^{t-} =_{df} \beta$ presented a novel negative argument to α at t

$S_{\beta\alpha}^{t} =_{df} \beta$ presented a novel negative or a novel positive argument to α at t

The following is a consequence of Laputa under the argumentation interpretation:

If $T_{\beta\alpha} > 0.5$, β presents a positive argument to α only if $C_{\beta}(p) \geq T_{\beta\alpha}$, and a negative argument only if $C_{\beta}(p) \leq 1 - T_{\beta\alpha}$.

Thus, if the threshold of assertion exceeds 0.5, then the inquirer will present an argument, whether it be positive or negative, if her confidence in the conclusion exceeds the threshold. A threshold of assertion exceeding 0.5 captures a sense in which deliberating agents are *sincere*. While this is surely the normal case, the model is general enough to allow for inquirers to be insincere, in the following sense:

If $T_{\beta\alpha} < 0.5$, β utters a positive argument to α only if $C_\beta(p) \leq T_{\beta\alpha}$, and negative argument only if $C_\beta(p) \geq 1 - T_{\beta\alpha}$.

In other words, a threshold of assertion below 0.5 is interpreted as a "liar threshold": the inquirer will give an argument for *p* only if her degree of belief in p is sufficiently low; and an argument for not-*p* only if her degree of belief in not-*p* is sufficiently low. Setting the threshold of assertion to a number below 0.5 can be used to model a kind of strategic communication, e.g., lying or acting as the "devil's advocate", in the sense of giving an argument for *p* (not-*p*) while personally believing *p* (not-*p*) to be false. Finally, if $T_{\beta\alpha} = 0.5$, β can utter a positive or a negative argument to α independently of what she believes, which is modeled by letting her pick what to say randomly.[3]

The source independence assumption states that inquirers treat other inquirers as giving independent information. Whether or not we choose the opinion disclosure or the argumentation interpretation of Laputa, assuming source independence has the effect of disconnecting inquirers from reality after a few deliberative rounds. The reason is that inquirers, when updating their credences in *p*, will take into account not only the result of their own inquiries but also the information coming from other inquirers, whether that information is interpreted as disclosed opinions or novel arguments. This will lead to the credences of inquirers becoming, with time, increasingly dependent. After a while, positive (negative) reports coming from other inquirers cannot be taken anymore as independent indications that *p* (not-*p*) is true, and yet the listening inquirers in Laputa will treat them as such. No reason has been presented, however, indicating that source independence systematically distorts simulation results in any particular direction. And arguably, source independence is psychologically realistic as a default strategy: real inquirers have a tendency to assume source independence in the absence of concrete reasons to think that sources are not independent. Keeping in mind the considerable simplifying effects source independence has on the entire model, we are therefore justified in accepting it as a highly useful idealization.

5. Do Bayesian inquirers polarize?

[3] The original idea behind Laputa was to simulate communication based on inquiry. A possible drawback with the argumentation interpretation is that it decouples inquiry from communication. The existence of arguments is not brought in relation to the result of inquiry, and whether or not an inquirer possesses an argument is not represented in the model. We have been experimenting with a version of the program in which communication is possible only if new inquiry has taken place. Preliminary simulations suggest that this modification does not have any significant statistical effect on simulation outcome.

In early work in social psychology it was observed that group decisions are sometimes riskier than the previous private decisions of the group's members.[4] This observation paved the way for numerous studies showing that *risky shift* is a pervasive phenomenon but also that on certain decisions groups are actually more cautious that their members. Both risky and cautious shifts are special cases of a group-induced attitude *polarization* (e.g. Moscovici and Zavalloni, 1969). Group polarization is said to occur when "an initial tendency of individual group members toward a given direction is enhanced following group discussion" (Isengren, 1986, p. 1141) so that "members of a deliberating group predictably move toward a more extreme point in the direction indicated by the members' predeliberation tendencies" (Sunstein, 2002, p. 176, italics removed). Thus, a group of moderately profeminist women will be more strongly profeminist following group discussion (Myers, 1975).

Given that "[g]roup polarization is among the most robust patterns found in deliberating bodies" (Sunstein, 2002, p. 177), we can use polarization as a test of empirical adequacy that any reasonably realistic model of group deliberation should satisfy. In this section we test whether inquirers in Laputa polarize under what would appear to be normal circumstances characterized by (i) some prior trust in the reliability of the others, (ii) an inclination to give arguments only if the conclusion is perceived to be more likely to be true than false, and (iii) an admission to talk in the absence of a high degree of credence in the conclusion. Figure 3 shows the exact parameter settings in the batch window of Laputa.

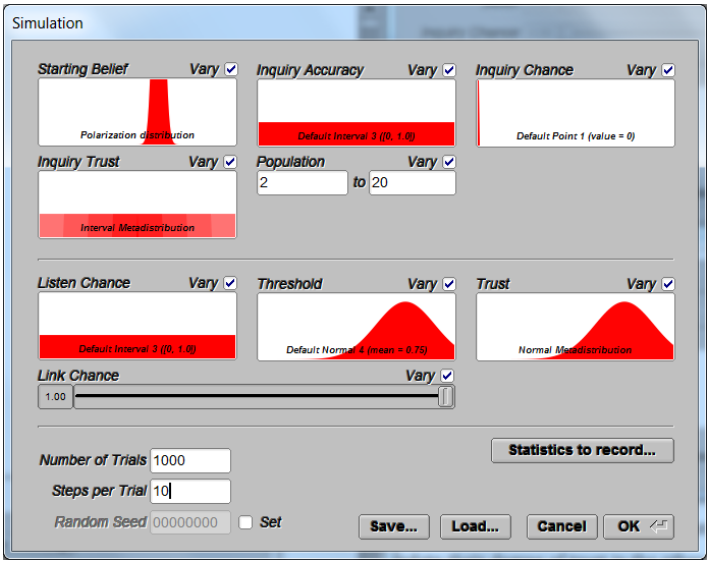

Figure 3

[4] Isengren (1986) credits an unpublished master thesis by James Stoner with this discovery (Stoner 1991).

It was assumed that the inquirers engage in a "closed room" debate without undertaking any inquiry whilst deliberating. Hence, the inquiry chance parameter was set to 0 and the link chance to 1, making every announcement public within the group. The threshold of assertion was taken to be normally distributed around 0.75. The social trust parameter (credence in the reliability of others) was assumed to be normally distributed in the area above 0.5. Finally, the initial degree of belief (credence) in *p* was taken to be positive and normally distributed just above 0.5. Laputa was then instructed to generate 1,000 networks ("trials") satisfying these constraints, allowing each network to evolve 10 steps. The result is depicted in figure 4.

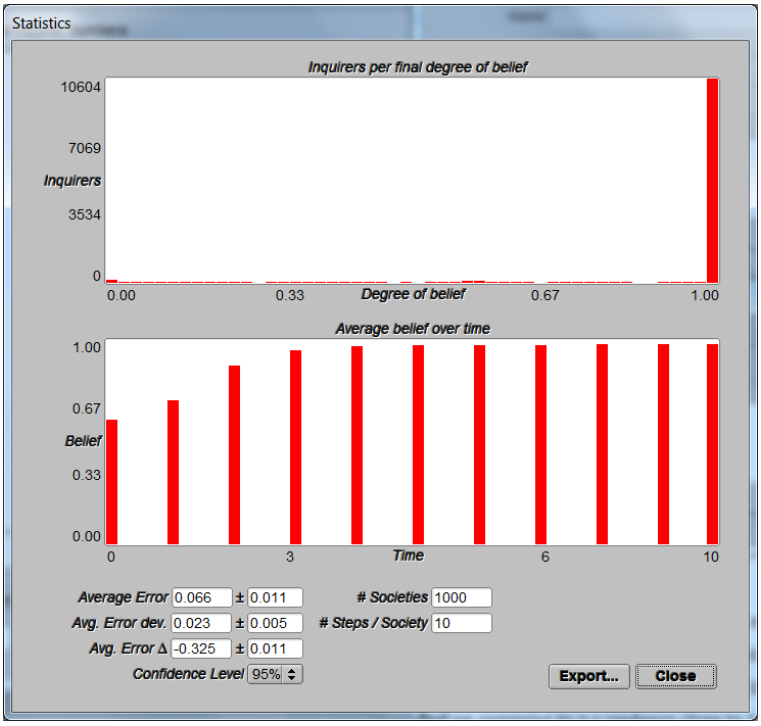

Figure 4

The lower diagram of figure 4 shows the evolution of the average credence in *p* over time. As we see, after a view steps the average credence in *p* converged to a value slightly below 1. The upper diagram of figure 4 shows the number of inquirers per final credence in *p* after 10 deliberative rounds. Virtually all inquirers ended up assigning *p* a credence close to 1. These observations confirm our prediction: inquirers in Laputa polarize in the sense that if every inquirer is initially inclined to believe *p*, however cautiously, they will still believe *p* after deliberation, only much more strongly. The effect is the same *mutatis mutandis*, if the inquirers initially favor not-*p* rather than *p*, in which case they will end up believing not-*p* more strongly.

We recall that inquirers in Laputa update their degree of trust in the other inquirers dynamically, although we have not detailed the mechanisms behind trust in this paper (see

Angere forthcoming). Intuitively, we would expect polarization with regard to the proposition at stake to be accompanied by increased mutual trust among the inquiring agents. This is indeed what happens in Laputa. This effect was studied for a small network of only two inquirers under circumstances similar to those in figure 3. More precisely, communication chances for inquirer 1 and inquirer 2 were set to 0.94 and 0.88, respectively; and the threshold for the links outgoing from inquirer 1 and outgoing from inquirer 2 were set to 0.66 and 0.67, respectively. Figure 5 shows the result.

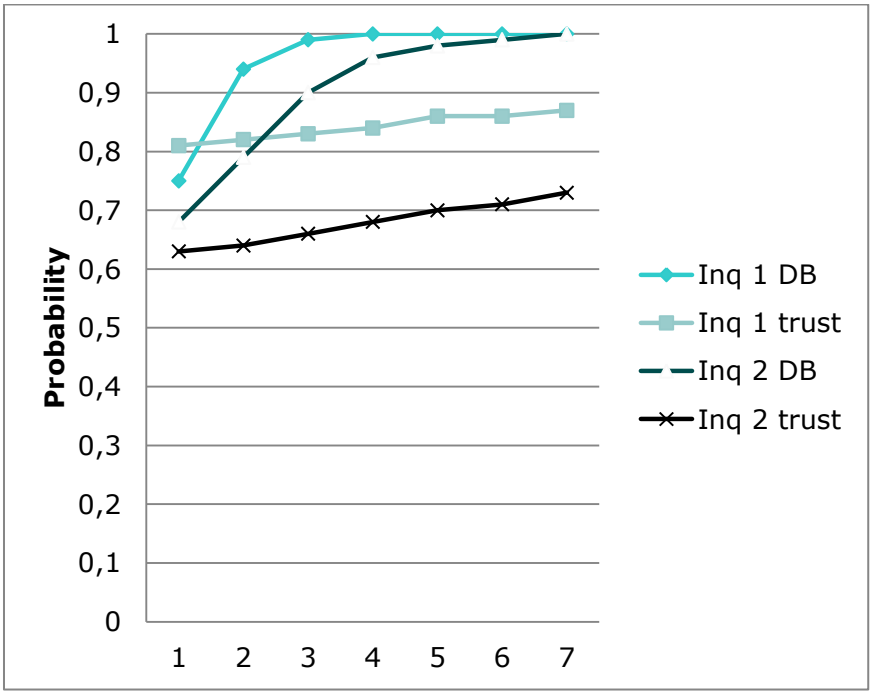

Figure 5

The horizontal axis shows time. The vertical axis displays the relevant credences. We see that as the inquirers polarize with regard to their credence, or degree of belief (DB), in *p*, they become increasingly more trusting vis-à-vis each other.

These results are easily explained given what we know about the underlying Bayesian model. If the inquirers are initially inclined towards *p* and some have a threshold of assertion allowing them to communicate, the latter will give novel arguments in favor of *p*. These arguments will be taken into account by the listening inquirers in the manner previously described, leading them to adjust their credence in *p*, as well as their trust in the source, upwards (see cell (a) in table 1). With time, an increasing number of inquirers will find their credence in *p* exceed their threshold of assertion, encouraging them to give further arguments for *p*. This will push credences in *p* still

further in the positive direction, and so on. At the same time, the growing sense of being confirmed by the others will lead to increased mutual trust among the inquirers, adding further momentum to their polarization. Moreover, this also shows that, in normal cases, polarization on the belief level is accompanied, in a sense, by polarization on the trust level: the initially shared attitude of trust is reinforced as the effect of deliberation.

Our study raises the further question what happens in more unusual cases, e.g. when people do not trust each other or they "lie" in the slightly technical sense of giving arguments for a conclusion they do not believe in. There are three cases to consider: people trust but lie, people distrust but tell the truth, people distrust and lie. Using our simulation program, we tested these three cases while keeping all the other assumptions intact. The results are summarized in table 2.

	Trust	**Distrust**
Truth-telling	Polarization	Divergence
Lying	Divergence	Polarization

Table 2

As we see, there are two situations that lead to polarization, as always in the sense that like-minded people are strengthened in their initial convictions as the effect of deliberation. One is the normal situation which we studied in the previous section, i.e. when people trust other people and do so for good reasons because the others are in fact trustworthy. The other is when people distrust others, again for good reasons because the others are in fact untrustworthy. These two cases exemplify what we might call situations of *social calibration*: people's attitudes towards other people adequately reflect the actual trustworthiness of the latter. In the two remaining cases, in which there is lack of social calibration, we typically get a divided society: one camp believing the truth and the other camp believing the falsehood with the members of one camp distrusting the members of the other.[5]

[5] At the end of a batch simulation, Laputa outputs the distribution of average final degrees of belief for all inquirers in all societies that were considered. For an example see the upper diagram of Figure 4. Laputa, as it stands, does not output the distribution of final degrees of belief for particular societies. Hence, we cannot conclude that societies that are not socially calibrated will divide from the data that we get from Laputa while in batch mode. However, during a batch simulation Laputa randomly selects societies for

We will close this section by studying an example of how a society consisting of serious (truth-telling) inquirers initially inclined to believe the same thing can still end up divided on the issue as the effect of a lack of social calibration. We will study a simple society consisting of only two inquirers: Inquirer 1 (Inq 1) and Inquirer 2 (Inq 2). We set listen chance for Inquirer 1 to 0.94 and for Inquirer 2 to 0.88, and the threshold for both inquirers to 0.58. We choose a normally distributed trust function for both inquirers with expected value 0.38. Figure 6 shows how the inquirers degree of belief (DB) in p, and the expected value of their trust functions, change with time.

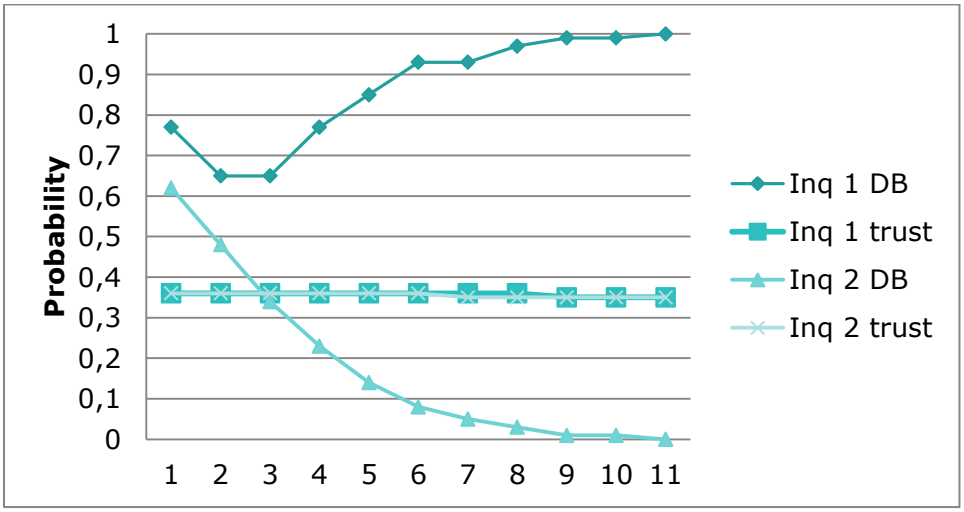

Figure 6

We see that after some fluctuations the general trend is that one inquirer will start believing *p* while the other will start believing not-*p*.

Laputa allows us to inspect the relevant parameters in a step-wise fashion to see what causes this result. This reveals that the following transpires:

1. Both inquirers initially give arguments for *p* because their DB in *p* is above the threshold of assertion.
2. Since they distrust each other, they will take each other's arguments as evidence for not-*p* and lower their DB in *p* (see cell (g) in table 1).
3. Inquirer 1 still has a DB in *p* which is above the threshold, and so she gives an argument for *p*.
4. Given her distrust in Inquirer 1, Inquirer 2 becomes rather confident that not-*p*, so she gives an argument for not-*p* (cell (g) in table 1).

---

visual representation on the computer screen. That visual information was used as additional data when concluding that a divided society results under the conditions given in table 2.

5. Given her distrust in Inquirer 2, this is taken by Inquirer 1 to be evidence for the opposite, namely *p* (cell (i) in table 1).
6. By the same token, Inquirer 1 will continue to argue for *p* while Inquirer 2 will continue to argue for not-*p*, and they will become ever more confident in the conclusions of their arguments.
7. Eventually Inquirer 1 will become certain of *p* and Inquirer 2 certain of not-*p*.
8. Meanwhile they will continuously downgrade their degree of trust still further because, as they see it, they repeatedly receive surprising messages from a distrusted source (see cell (i) in table 1).

We note that while divergence occurs with respect to credence in p, polarization occurs with respect to trust: the inquirers initially distrusted each other and this initial tendency is reinforced as the effect of deliberation.

6. Conclusion and discussion

I have argued that the original interpretation of Laputa as a model of opinion disclosure is somewhat problematic due to a problem of repetition. Instead, I proposed an interpretation according to which inquirers are exchanging novel arguments for or against a target proposition. I went on to show that the model exhibits polarization much like real argumentative bodies. Inquirers in Laputa, if initially disposed to believe in a given proposition, will see their credences in that proposition increase as a result of group deliberation. This lends additional credibility to the model as a reasonably realistic representation of the phenomena in question. We also studied conditions under which inquirers diverge in their opinions. To the extent that Bayesian reasoning is normatively correct, the perhaps most surprising, and disturbing, results of this study is that polarization and divergence are not necessarily the result of mere irrational "group think" but that even ideally rational inquirers will predictably polarize or diverge under realistic conditions. It remains to compare the present theory with the influential Persuasive Argument Theory (PAT) which also predicts polarization.[6]

According to PAT an individual's position on an issue is a function of the number and persuasiveness of pro and con arguments that the person recalls from memory when formulating his or her own position. Thus in assessing the guilt or innocence of an accused in trial, jurors come to predeliberation decisions on the basis of the relative number and persuasiveness of arguments favoring guilt or innocence. Group deliberation will cause an individual to shift in a given direction to the extent that the discussion exposes that individual to persuasive arguments favoring that direction rather than to arguments favoring the opposite

[6] The following account is based on the overview in Isenberg (1986), pp. 1145-1148.

direction. How persuasive an argument is to a given individual is determined by the validity and novelty of the argument. One factor, among several, affecting perceived validity is the extent to which the argument fits into the person's previous views. Novelty has to do with how new and unusual the argument is to the person in question. Everything else equal, a novel argument has a greater persuasive force than a commonplace argument.

Laputa, as I have proposed to interpret it, is clearly in the spirit of PAT. Thus, Laputa is also based on the assumption that the persuasive effect of an argument depends essentially on two factors: its perceived validity (including the trustworthiness of the presenter) and novelty. There are also differences. For instance, Laputa is more specific than PAT in assuming that individual inquirers update their degrees of belief in a particular way, namely that dictated by Bayesianism. PAT as such does not postulate any more specific updating mechanism, let alone a Bayesian one. Laputa assumes, in addition, that individuals' degrees of trust are dynamically updated in a Bayesian fashion.

Furthermore, inquirers in Laputa engaging in group deliberation update their credences in a piecemeal or sequential fashion. The presentation of a novel argument, or collection of arguments, will normally affect the receiving inquirer's credence in the conclusion. As PAT is normally formulated, inquirers are supposed to collect in memory all the arguments they are presented with during group deliberation, postponing their own verdict on the matter until deliberation has come to an end. When the deliberation has ended the inquirer takes a stand on the basis of a holistic assessment of the number and merits of the pro and con arguments retained in memory. This "holistic" aspect of PAT is not unproblematic in the light of experiments indicating that the order in which arguments are presented will affect the conclusion reached. Thus, Kaplan and Miller (1977) found that subjects tend to recall persuasive arguments that they had been exposed to most recently rather than the ones they had been exposed to first.

While there may be doubts about some of the details of PAT, there are many experimental studies pointing to its broad empirical adequacy. It is reasonable to suppose that a fair number of these studies will give (indirect) support for Laputa under the argumentation interpretation considering the fact that the latter is, by and large, subsumable under the former. A more careful assessment of this claim which has the status of a reasonable conjecture is, however, outside the scope of the present article.[7]

Appendix

[7] Acknowledgement: I am grateful to Staffan Angere and Stephan Hartmann for their comments on previous versions of this paper.

To illustrate the role played by the condition of source independence, we consider the case of one inquirer α receiving, at time t, positives messages from two sources, σ_1 and σ_2. By (Cond),

$$C_\alpha^{t+1}(p) = C_\alpha^t(p \mid S_{\sigma_1\alpha}^{t+} \wedge S_{\sigma_2\alpha}^{t+})$$

$$= \frac{C_\alpha^t(p) C_\alpha^t(S_{\sigma_1\alpha}^{t+} \wedge S_{\sigma_2\alpha}^{t+} \mid p)}{C_\alpha^t(S_{\sigma_1\alpha}^{t+} \wedge S_{\sigma_2\alpha}^{t+})} \qquad \text{(Bayes' theorem)}$$

$$= \frac{C_\alpha^t(p) C_\alpha^t(S_{\sigma_1\alpha}^{t+} \wedge S_{\sigma_2\alpha}^{t+} \mid p)}{C_\alpha^t(S_{\sigma_1\alpha}^{t+} \wedge S_{\sigma_2\alpha}^{t+} \mid p) C_\alpha^t(p) + C_\alpha^t(S_{\sigma_1\alpha}^{t+} \wedge S_{\sigma_2\alpha}^{t+} \mid \neg p) C_\alpha^t(\neg p)} \qquad \text{(Total probability)}$$

$$= \frac{C_\alpha^t(p) C_\alpha^t(S_{\sigma_1\alpha}^{t+} \mid p) C_\alpha^t S(_{\sigma_2\alpha}^{t+} \mid p)}{C_\alpha^t(S_{\sigma_1\alpha}^{t+} \mid p) C_\alpha^t(S_{\sigma_2\alpha}^{t+} \mid p) C_\alpha^t(p) + C_\alpha^t(S_{\sigma_1\alpha}^{t+} \mid \neg p) C_\alpha^t(S_{\sigma_2\alpha}^{t+} \mid \neg p) C_\alpha^t(\neg p)} \qquad \text{(Source independence)}$$

$$= \frac{C_\alpha^t(p) \langle \tau_{\sigma_1\alpha}^t \rangle \langle \tau_{\sigma_2\alpha}^t \rangle}{C_\alpha^t(p) \langle \tau_{\sigma_1\alpha}^t \rangle \langle \tau_{\sigma_2\alpha}^t \rangle + C_\alpha^t(\neg p) \langle \bar{\tau}_{\sigma_1\alpha}^t \rangle \langle \bar{\tau}_{\sigma_2\alpha}^t \rangle} \qquad \text{(By (*) and (**))}$$

This means that we only need three pieces of information in order to compute α's posterior credence in p: α's prior credence in p, the expected value of α's trust function for σ_1 and for σ_2. Supposing these values to be 0.8, 0.7 and 0.9, respectively, we get a 0.99 posterior credence in p.

The example can be generalized as follows:

Theorem 1: Suppose that α at t receives messages from exactly n sources σ_1, …, σ_n, and that all messages are positive. Then

$$C_\alpha^{t+1}(p) = \frac{C_\alpha^t(p) \prod_{i=1}^n \langle \tau_{\sigma_i\alpha}^t \rangle}{C_\alpha^t(p) \prod_{i=1}^n \langle \tau_{\sigma_i\alpha}^t \rangle + C_\alpha^t(\neg p) \prod_{i=1}^n \langle \hat{\tau}_{\sigma_i\alpha}^t \rangle}$$

Proof: Left to the reader.

We can generalize this still further.

Theorem 2: Suppose that α at t receives messages from exactly n sources σ_1, …, σ_n, Let Pos_α^t be the set of all indices of sources giving positive messages, and $Negs_\alpha^t$ be the set of all indices of sources giving negative messages. Then

$$C_\alpha^{t+1}(p) = \frac{C_\alpha^t(p) \prod_{i \in Pos_\alpha^t}^n \langle \tau_{\sigma_i\alpha}^t \rangle \prod_{i \in Neg_\alpha^t}^n \langle \hat{\tau}_{\sigma_i\alpha}^t \rangle}{C_\alpha^t(p) \prod_{i \in Pos_\alpha^t}^n \langle \tau_{\sigma_i\alpha}^t \rangle \prod_{i \in Neg_\alpha^t}^n \langle \hat{\tau}_{\sigma_i\alpha}^t \rangle + C_\alpha^t(\neg p) \prod_{i \in Pos_\alpha^t}^n \langle \hat{\tau}_{\sigma_i\alpha}^t \rangle \prod_{i \in Neg_\alpha^t}^n \langle \tau_{\sigma_i\alpha}^t \rangle}$$

Proof: Left to the reader.

Corollary 1: Suppose that α at t receives messages from exactly n sources σ_1, …, σ_n, for an even $n > 0$, that $\langle \tau^t_{\sigma_i \alpha} \rangle = \langle \tau^t_{\sigma_j \alpha} \rangle$, and that there is an equal number of positive and negative messages. Then $C^{t+1}_\alpha(p) = C^t_\alpha(p)$.
Proof: Follows from theorem 2.

References

Angere, S. (forthcoming), "Knowledge in a Social Network", submitted paper.
Goldman, A. I. (1999), *Knowledge in a Social World*, Clarendon Press, Oxford.
Grice, P. (1975), "Logic and Conversation", in Cole, P. and Morgan, J. (eds.), *Syntax and Semantics, 3: Speech Acts*, New York: Academic Press: pp. 41–58.
Hegselmann, R., and Krause, U. (2006), "Truth and Cognitive Division of Labour: First Steps Towards a Computer-Aided Social Epistemology", *Journal of Artificial Societies and Social Simulation* 9 (3).
Isenberg, D. (1986), "Group Polarization: A Critical Review and Meta-Analysis", *Journal of Personality and Social Psychology* 50 (6): 1141-1151.
Kaplan, M. F. (1977), "Discussion Polarization Effects in a Modified Jury Decision Paradigm: Informational Influences", *Sociometry* 40: 262-271.
Kaplan, M. F., and Miller, C. E. (1977), "Judgments and Group Discussion: Effect of Presentation and Memory Factors on Polarization", *Sociometry* 40: 337-343.
Myers, D. G. (1975), "Discussion-induced Attitude Polarization", *Human Relations* 28: 699-714.
Moscovici, S., and Zavalloni, M. (1969), "The Group as a Polarizer of Attitudes", *Journal of Personality and Social Psychology* 12: I25-135.
Olsson, E. J. (2005), *Against Coherence: Truth, Probability, and Justification*, Oxford University Press.
Olsson, E. J. (2011), "A Simulation Approach to Veritistic Social Epistemology", *Episteme*, in press.
Olsson, E. J., and Vallinder, A, (in press), "Norms of Assertion and Communication in Social Networks", *Synthese*.
Vallinder, A., and Olsson, E. J. (in press a)), "Trust and the Value of Overconfidence: A Bayesian Perspective on Social Network Communication", *Synthese*.
Valinder, A., and Olsson, E. J. (in press b)), "Does Computer Simulation Support the Argument from Disagreement?, *Synthese*.

Stoner, J. A. F. (1961), *A Comparison of Individual and Group Decisions Involving Risk.* Unpublished master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Sunstein, C. R. (2002), "The Law of Group Polarization", *The Journal of Political Philosophy* 10 (2): 175-195.

Vinokur. A., and Burnstein, E. (I978), "Novel Argumentation and Attitude Change: The Case of Polarization Following Group Discussion", *European Journal of Social Psychology* 8*:* 335-348.

Zollman, K. J. (2007), "The Communication Structure of Epistemic Communities", *Philosophy of Science* 74 (5): 574-587.