When Learning Ends: Do Evidence Processing Biases Hinder Consensus-Building?

Introduction

Idealizations of democratic and market processes postulate that the widespread exchange of information is a net social benefit. In particular, when evidence about a phenomena is distributed across a community, free-flow exchange of information should be beneficial: information-sharing allows individuals to aggregate evidence that otherwise would be widely distributed and inaccessible.

Several agent-based models of information-exchange have been developed. These models envision a strongly connected network of agents that share information and update their beliefs over time. Researchers have focused on identifying whether agents in a given model converge on a single belief about the world in the long-run, and examined which factors impair the formation of long-run consensus.

Previous examinations of these models have two shortcomings. First, the researchers almost invariably examine consensus formation in the very long-run, by taking the limit as time goes to infinity of the mathematical models they've developed.[1] As [1] notes, the forms of social learning that we're concerned with occurs on finite, and in some cases, relatively short timeframes; what happens in the long-run may not be terribly important. The second shortcoming is that most of the models that have been developed do not account for the foibles of human information processing. I take this to be a weakness since most researchers are interested in claims about human societies (and not, say, autonomous systems).

This paper addresses these two unexplored shortcomings. I explore short timeframes by relying on simulations and I examine the impact of biased evidence processing on consensus formation.

Prior Work

This paper draws on two research literatures: research on social learning, and research examining how humans evaluate evidence.

Two concerns have dominated the social learning literature: 1) under what conditions will a community of agents come to agreement on their beliefs about an arbitrary world-state and 2) if the community does reach a consensus, how close will the group consensus be to the true

[1] [8], which examines a somewhat different phenomena involving diversity, is an exception.

state of the world? All models demonstrate the interplay between consensus and accuracy. One simple yet striking model is Banerjee's urn model, where agents sequentially receive a private signal and witness the actions of all agents who have acted before them. Agents learn about the state of the world through private and public signals. If agents weigh all signals equally, the group will quickly form a consensus (and ignore private signals). But under the right conditions, there is an uncomfortably high probability that the consensus will be false.

As social networks have become more prominent, researchers have developed models where the agents are embedded in a social network, either implicitly or explicitly. Each agent receives signals from a subset of other agents, and applies an update rule involving these signals and the agent's prior belief at each time step to generate the agent's new belief or next action. The idealized update rule is Bayesian updating, but various others have been proposed, in part because Bayesian updating requires extensive knowledge that social agents in real-world settings do not have. Under the right conditions, these models predict that information-sharing will lead to consensus in the long-run.

What, then, explains a lack of consensus on many public issues despite widespread information-sharing? Researchers have proposed several ideas. One explanation is malevolent agents. Acemoglu and Ozdaglar demonstrate that sufficiently well-connected agents intent on misinforming the community can prevent the community as a whole from reaching consensus. [1] For example, a large media organization intent on misleading the public could prevent public consensus from forming. Another idea along these same lines is Acemoglu and Ozdaglar’s “forceful” agents. [2] These are agents who, because of irrationality or ulterior motives, refuse to update their beliefs about the state of the world (which creates misinformation).

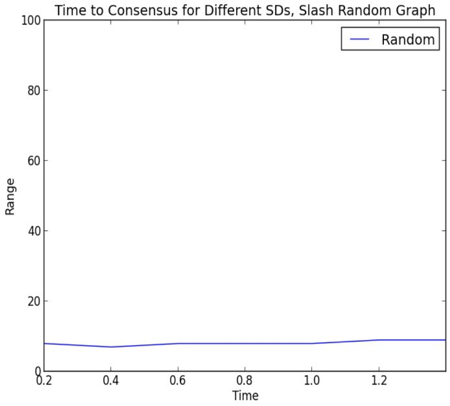
It’s unsurprising that a sufficient number of malevolent agents can sabotage consensus-building. Can well-meaning agents also fail to generate consensus through information-sharing? My suspicion is that simple biases in evidence processing can explain the lack of consensus.[2] I apply two psychological biases to the agent-based models that have been developed by others to examine the extent to which these biases would impair consensus-formation.

Psychological literature of the past fifty years have demonstrated that the human evidence assimilation is full of quirks and biases, at least when compared to ideal, Bayesian updating. I examine the effect of two theories of human evidence processing: biased assimilation and attitude polarization, which I discuss below.

Model and Method

[2] I'm not discounting the idea that malevolent or irrational agents exist in the social world; just that a community without such agents could or perhaps would fail to reach consensus under the conditions I examine.

I adopt the naive learning model outlined in [3]. This model has several nice characteristics. It's simple to implement; like Bayesian updating, in the long-run, the community will reach consensus; unlike Bayesian updating, it does not require agents to have extensive knowledge about the conditional probabilities of signals given states nor a priori probabilities of true states. I take two directed graphs: a real graph exhibiting common network characteristics (high clustering coefficient, low diameter, power law degree distributions) and a random graph generated with Snap with the same number of nodes and edges.[3] I deleted nodes that had zero in-degree (as such nodes would never update their beliefs). Both graphs are weakly connected (which is necessary for these simulations to work). I generate the true state uniformly at random on the interval [0,1]. To reflect the similarity of beliefs among neighbors, I seed beliefs in a clustered way. Working from a distribution with μ equal to the true state and standard deviation σ, agents in the same neighborhood draw their prior beliefs from a distribution with μ equal to a single sample of the above distribution and standard deviation σ.[4] For all of the following simulations, I fix $\sigma = .25$. Varying σ does not effect the time it takes to reach consensus, when consensus is defined as when the range of beliefs hit $k\sigma$ (Right: Varying σ does not alter the time to consensus, where consensus = $\sigma/5$.



Thus, at time t_0, each agent has a belief about the true state of the world. At each following time-step, agents communicate their beliefs about the true state of the world truthfully to their neighbors.

The Effect of Graph Structure on Consensus Formation

Real social networks are highly clustered. We would expect, all other things being equal, that a highly clustered network would take longer to reach consensus than a lower clustered network, as beliefs within clusters will quickly converge, but the transmission of information

[3] The real graph is the February 2009 Slashdot social network, available at http://snap.stanford.edu/data/soc-Slashdot0902.html (~ 80,000 nodes, ~ 950,000 edges). Another social network was also identified, but analysis on that is incomplete.

[4] This was done by choosing nodes at random and creating a neighborhood from the chosen node and its neighbors. Existing clustering algorithms, even fast ones, were simply taking too long to be practical.

across bridges in the network will be slow. This is particularly true if neighborhoods begin with geographically similar priors.

Real social networks also have power-law degree distributions. Under the naïve learning model, highly influential nodes have a disproportionate effect on the final consensus reached. [3] At every time-step, the belief of the influential node influences many other agents. In other words, if an influential node happens to have an extreme belief, her influence will pull the consensus estimate away from the true state of the world. Influential nodes vary by how much they are influenced, however. Agents that influence the influential node also play a large role in pulling the consensus estimate. Agents with high in-degree and high out-degree are distributors of relatively accurate information – they help make the consensus of the group accurate. Agents with low in-degree but high out-degree in some sense oversell the information that they have.

One of the questions this paper explores is whether the graph structure found in real-world networks exacerbates or inhibits the psychological biases I implement.

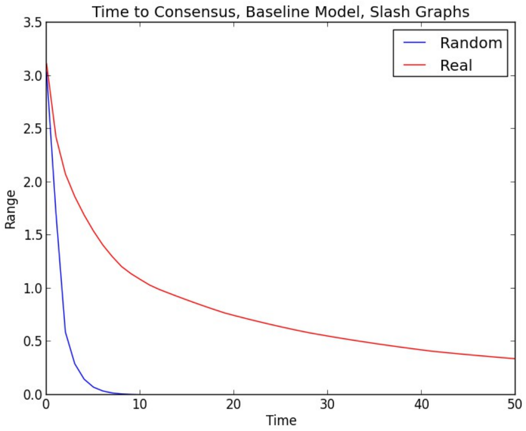
Baseline model

So far, this model is in accordance with the models discussed in [1-4]. At each time step, information flows from each agent *i* to the set of *i*'s outgoing links. For the baseline model, the new belief of agent *i* is the average of the signals she has received from her neighbors and her prior belief. Equivalently, the belief of agent *i* at time *t* is as follows:

$$b_i(t+1)=\frac{\sum_{k=1}^{k=n} b_k(t)+b_i(t)}{n+1}$$

Where *n* is the number of signals that agent received. Equivalently, *n* – 1 is the number of in-links the agent has. Under this model, all agents will converge on a belief about the true state of the world. [3]

I first develop baseline results, using this simple model before implementing any extensions. The results are below.



Each time step represents a re-evaluation of available evidence by the agents and an update to existing agent beliefs. The relationship between time-step and real-world time is context-dependent. Interacting players in a fast-paced game might be exchanging information relatively quickly; in other cases agents might update their beliefs relatively rarely (Left: time to consensus for a real network and a random network with σ = .25, average of iterations).

Both networks approach consensus, but the random network reaches consensus in about 8 time-steps. The real network still has a belief range close to 2σ even after 50 iterations. This suggests that the structure of real networks inhibits consensus formation, at least under the naïve learning model outlined in [3]. More investigation is warranted here, but I suspect the presence of very high degree nodes play a role in the result. Take the 1000 agents with the largest out-degree. With a purely random seeding model (where each node gets a prior belief drawn from a normal distribution with mean μ and standard deviation σ), we would expect 50 agents to have prior beliefs more than 2σ away from the mean. These highly influential nodes prevent consensus from forming too quickly.

Biased Assimilation

I implement two extensions of the baseline model. The first extension of the baseline model is the implementation of biased assimilation. Biased assimilation involves favoring evidence that is closer to your initial views over equivalently reliable evidence that departs from your views. Instead of equally weighing each signal that an agent receives, as in the baseline model, under biased assimilation agents differentially weight the signals that they receive. Under the baseline model, the weight of each signal was simply $\frac{1}{n}$. Biased assimilation alters these weights (under-valuing distant signals and consequently over-valuing near signals).

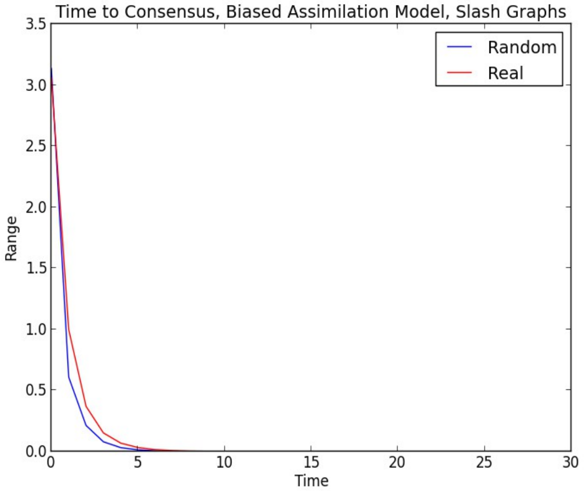
I interpret assimilation bias as follows. From the perspective of agent *i*, the weight of neighbor *k*'s signal under biased assimilation is multiplied by $\frac{r-d_{ik}}{r}$, where $d_{ik}=|b_i-b_k|$

and b_n is the belief of node n. *r* represents the range of beliefs in the network. Under this interpretation, weights decrease linearly as a function *d*.

$$w_k = \frac{r - d_{ik}}{nr}$$

The respective weights of agent *i*'s neighbors are then renormalized to sum to 1.

Another way of interpreting biased assimilation is that the signal an agent receives not only tells the agent about the potential state of the world, but also tells the agent something about the reliability of the source of the signal. A signal that is far away from an agent's prior belief is unlikely to be true, from the agent's perspective. Other researchers have incorporated agents' estimates of precision into their models. This requires weighing each signal that an agent receives by a value indicating how much the agent trusts that signal to be reliable. The only difference is that these weights are not fixed.



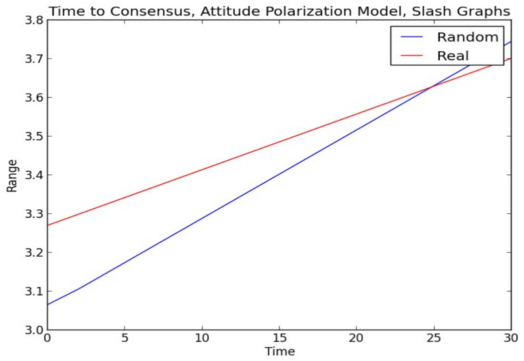
The results of simulations using this model are curious (Left: time to consensus in a biased assimilation model, same networks as above). Both real and random networks fall to consensus levels quite quickly. I had anticipated that biased assimilation would lead to delayed onset of consensus, but the opposite appears to be the case. The difference between the baseline and the biased assimilation models for the real graph is strange, but despite lots of tweaking, these are the results. Note that biased assimilation does not seem to alter how quickly random graphs reach consensus.

Attitude Polarization

The second extension of the baseline model is the implementation of attitude polarization. At least some psychological research indicates that human evidence processing involves an effect beyond biased assimilation. [6,7] Attitude polarization occurs when contrary evidence causes the subject to reinforce her initial belief. For example, assume you have a strong belief in the deterrence value of capital punishment. You read an article that strongly indicates that capital punishment does not deter crime. Attitude polarization occurs when you walk away from that encounter more convinced of the deterrence value of capital punishment than you were initially. Evidence for the existence of attitude polarization is mixed, but the bias appears to disproportionately affect those who hold extreme views.

Accordingly, I implement attitude polarization in the following way. First, I distinguish agents who have extreme beliefs as those who have beliefs more than β standard deviations away from the mean belief of the group. At t_0, these will be agents whose prior beliefs are more than β standard deviations away from the true state (as priors are seeded using the true state as the mean of a normal distribution). When an agent who has an extreme belief encounters a signal that is more than 2β standard deviations away from his current belief, he updates his belief as if he had received a signal in the opposite direction of the signal, by some small margin, k. In other words, if $b_i > \mu_b + \beta\sigma \wedge |b_i - b_j| > 2\beta\sigma$, the signal coming from node j to i is $s_j = k$. Similarly, if $b_i > \mu_b - \beta\sigma \wedge |b_i - b_j| > 2\beta\sigma$, the signal coming from node j to i is $s_j = -k$. This interpretation is simplistic, but accords with psychological evidence indicating that only individuals who hold extreme views are vulnerable to attitude polarization. Although it may seem that the larger $|b_i - b_j|$ is, the larger k, the polarization, should be, we have little evidence quantifying exactly what the relationship should be.

Note under this interpretation of attitude polarization, if two nodes' beliefs begin to move apart they will continue to move apart (given an unchanging graph). Therefore, whether a given community of agents will reach consensus under attitude polarization only rarely depends on k. First, two connected nodes must qualify for attitude polarization somewhere in the graph. The chance of this depends on network size, the distribution of prior beliefs, and β. For example, if prior beliefs are distributed normally, .3% of agents will have beliefs greater than (or less than) 3σ (β = 3). As graph size grows large, the raw number of agents who have “extreme” beliefs increases, and, particularly if the number of edges increases faster than the number of nodes (i.e., densification occurs), the chance that two nodes will qualify for attitude polarization is very high. Of course, if no nodes qualify, attitude polarization simply doesn't occur (Left: attitude polarization implemented on a real and a random graph with β = 2.5 and k = .05; beliefs diverge).



Second, the local network structure needs to reinforce the difference between nodes *i* and *j*. If *i* and *j* are both listening to each other, in the absence of any other moderate signals, *i* and *j* will diverge. If, however, they both listen to a third node l that is feeding the mean belief of the network, the beliefs of nodes *i* and *j* may converge in spite of the affects of attitude polarization (if $k > w_l s_l$, then beliefs will still diverge). Simulations that vary β confirm that once β becomes large enough (around 5) these graphs converge again.

Conclusions

I've discussed the use of simulations to explore the formation of consensus in communities of networked agents. There is some evidence that knowledge of graph structure is important in understanding how consensus will form during meaningful timeframes. It does not appear, however, that evidence processing biases differentially affect networks with different graph structures. Networks will still reach consensus in the presence of biased assimilation and generally will not reach consensus under attitude polarization.

The lack of consensus in real-world information exchanges can be explained by more than just bad or irrational actors: at least one evidence processing bias, attitude polarization, creates an ever widening range of beliefs, and, even under ideal conditions, consensus can take a significant amount of time to form. Future work will hopefully explore the anomalous results in the biased assimilation model and examine in depth which graph structures are responsible for the results I've discussed.

<u>References</u>

[1] Acemoglu, Daron & Ozdaglar, Asuman. Opinion Dynamics and Learning in Social Networks. 1(1) Dynamic Games and Applications, (2011).

[2] Acemoglu, Daron, Ozdaglar, Asuman, & ParandehGheibi, Ali. Spread of (Mis)Information in Social Networks. 70.2 Games and Economic Behavior 194-227 (2010).

[3] Golub, Benjamin and Jackson, Matthew O. Naive Learning in Social Networks and the Wisdom of Crowds. American Economic Journal Microeconomics (2010)

[4] DeMarzo, Peter M., Vayanos, Dimitri, Zwiebel, Jeffrey. Persuasion Bias, Social Influence, and Unidimensional Opinions. 118(3) The Quarterly Journal of Economics 909-968 (2003).

[5] MacCoun, Robert J., Biases in the Interpretation and Use of Research Results. 49 Annu. Rev. Psychol. 259-87 (1998)

[6] Lord, Charles G., Ross, Lee, & Lepper, Mark R. Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence. 37(11) Journ. Person. Soc. Psych. 2098-2109 (1979).

[7] Koehler, Jonathan J. The Influence of Prior Beliefs on Scientific Judgment of Evidence Quality. 56 Org. Behav. Hum. Dec. Processes 28-55 (1993).

[8] Hong, Lu, Page, Scott E., & Riolo, Maria. Individual Learning and Collective Intelligence. Working Paper, available at http://vserver1.cscs.lsa.umich.edu/~spage/PAPERS/LearningandDiversity.pdf (2011).