Measures of Cognitive Distance and Diversity.

Johannes Castner

July 31, 2014

Contents

Abstract

I use a model of human causal learning, *Causal Support* (Tenenbaum & Griffiths, 2001), to derive a meaningful measure of *Cognitive Distance*–the degree to which two people differ in their opinions about the workings of the world. Next, I amend this measure to quantify the notion of *Cognitive Diversity.* Cognitive Diversity measures the degree to which opinions vary within a human collective, such as a political party or a research department of a firm or university. Measures of Cognitive Distance and Diversity are important for theoretical and empirical work which aims to link Cognitive Diversity to collective Wisdom–an organization's success in recognizing structure in the universe of interest–in order to make robust collective decisions in uncertain environments.

1 Introduction

We need to unearth the mechanisms guiding transitions from wise to unwise collectives, if we care to achieve a science of sustainable development. This is so, because most decisions that are relevant to sustainable development–however we may define it–are at least indirectly influenced by human collectives: committees, congress, corporate boards etc. One important set of machanisms relates the diversity of individual minds within a collective to the wisdom with which it solves its problems (collective decision and collective action).

Decisions and problem solving, involve preferences, experienced data and beliefs. The diversity of preferences and information (experienced data) have been in focus for much of the work that considers the heterogeneity of actors in economics and political science. Here, I will focus on the diversity of belief systems. In teams of people who work closely together, ultimate preferences are often the same–outcomes that favor the success of the team are preferred–and relevant information is swiftly shared among all members. Thus we might regard information and preferences as homogenous. But people's minds have slowly been forged by diverse forces. Aside from inherant differences, they are subjected to thousands of interacting influences from within and without their social settings. Thus, we should not be surprised to find that people think differently from one another, even if they commonly observe the same long stream of data. Of course, collectives might also be diverse in their preferences and in addition their members may select information sources differently–because of preference variation over the set of information sources–which causes people's received information to vary. Climate change discussions at all scales (from local to global) for example, are likely to draw attendees with diverse goals and beliefs, who also expose themselves selectively to different information channels. To more clearly understand and articulate the benefits and costs of cognitive diversity–the diversity of mechanistic (causal) beliefs–in any of the above settings, we must have a way to quantify cognitive distance and cognitive diversity. This is what this paper is about.

Socially constructed measures of diversity–for example along ethnic, gender and religious dimensions–are highly sensitive to context and interpretation. Yet, they have become a common public agenda of firms and organizations. This diversity agenda is promoted for ethical and esthetic reasons or as a result of political pressures, while perhaps the most relevant form of diversity–for the success and robustness of collectives–is related to cognition and beliefs[1].

Analogous with definitions of bio-diversity (Weitzman, 1992) that are designed to represent or reflect the amount of genetic information that is available to the system–for greater functional diversity under a greater range of circumstances–an organization's cognitive diversity must be defined in such a way that it reflects the collective's current mental tool kit. However, the analogy with bio-diversity has its limits. Weitzman's diversity measure, for example, applies to collections of objects that can be filed into natural discrete categories such as genetic species or languages and these must be hierarchically related by inheritance. In the most ideal circumstance–the one for which Weitzman's measure has the most meaning–the species' inheritance structure forms a tree. Human minds, if they inherit from each other, can not be arranged meaningfully in such structures and there is no natural typology for human minds. Cognitive differences are often small and subtle and thus complex languages have evolved to communicate those subtleties. Thus, the spaces of mental models and their representations–Bayesian Networks–must be continuous spaces and not sets of types. But then we need a continuous measure of diversity.

Hansen and Sargent (2008) make use of a continuous distance measure–between models and some true data generating process–that is related to the cognitive distance measure that I present here. Both, Hansen and Sargent's distance measure–the Kullback–Leibler divergence–and the one I present here–the square-root of the Jensen Shannon Divergence–can be calculated over models from a broader class than the one discused by Hansen and Sargent. The measure I propose improves the measure that Hansen and Sargent use. Additionally, I discuss how Hansen and Sargent's class of models is related to Bayesian Networks. Bayesian Networks were recently suggested by Tenenbaum and Griffiths as general representations of human cognitive models, regarding some causal system (Tenenbaum & Griffiths, 2008). We must allow for cognitive models with complex causal structures if we want to build theories of robust collective decisions in complex contexts. Bayesian Networks are good candidates because they allow for almost arbitrary causal complexity.

I advance a cognitive distance metric which is the square root of a mea-

[1] Note however, that, depending on the context, socially constructed measures of diversity might be highly correlated with cognitive diversity, and thus they may serve as more applicable proxies of cognitive diversity than membership size which is often constrained. Hence, focusing on these often more easily apparent forms of diversity might be justified also on grounds of organizational efficacy and robustness, if and when cognitive diversity can be shown to have such benefits.

sure known as the Jensen-Shannon Divergence (henceforth JSD). The JSD is a symmetrized version of the distance measure between some true process and an individual's model which is used by Hansen and Sargent in their important work on robust decisions in economics (Hansen and Sargent, 2008). The JSD is a lesser known, yet important quantity in information theory. When we take the square root of the JSD we obtain a metric, satisfying all of the properties of metrics. The older and more widely known measure used by Hansen and Sargent–the Kullback–Leibler divergence–neither satisfies symmetry nor the triangle inequality and thus the square-root of the JSD is an improvement as a measure of distance. The way in which we intuitively understand distance, demands distance measures to satisfy at least symmetry–if an object is at distance d from some other object, the second object better be at that same distance d from the first object–and less importantly it should satisfy the triangle inequality.

With a good measure of cognitive diversity we may show analytically and empirically whether and under what conditions cognitive diversity leads to robust collective decisions[2]. We may do this by relating cognitive diversity to other measures, such as one measuring collective wisdom. In addition, we may also construct a normative theory regarding how collectives should make robust decisions and how in constructing collectives, managers should trade off diversity against other concerns.

To further explain why robustness would follow from diversity, some people's models of the world–to be made precise below–that had been inaccurate previously, might have more explanatory power as circumstances change, while others that once furnished the best explanations might fade in relevance. A system of interest might cycle through regimes so that it be better explained by chaining together a set of simple models with probabilistic transition rules between them than by picking one complicated model. For different people different past experiences are salient and they build models that best explain those salient experiences. For example, a stock trader who lost money in the last financial crisis, will craft an exposed theory that explains it. A person's prevailing biases might then be useful for the group at particular times.

Starting with Condorcet's jury theorem, (Marquis de Condorcet, 1785)[3],

[2] The benefits and costs of diversity must depend also on the organization's opinion aggregation scheme, just as the benefits of bio-diversity depend on the structure of the food web.

[3] Although since Waldron (1995) there has been widespread argreement among scholars that in *Politics* Aristotle had already espoused a theory of "The Wisdom of the Multitude", which implicitly was synonymous with a theory of the social benefits of Diversity,

there have been various arguments with different degrees of sophistication, that, depending also on the organizational scheme of the collective, many thinkers together will on average come to wiser conclusions than any individual could on her own. In general, in these arguments the implied reason is that the greater the number of people, the greater is the variance in insights–the cognitive heterogeneity and diversity is higher in larger groups, by assumption. The diversity of minds and not the sheer number of people is in truth believed to lead to better collective decisions, but this is not how it is formulated in the theories[4]. For a thorough discussion and overview of the literatures on and related to collective wisdom, distributed intelligence and cognition or the wisdom of crowds, see "Collective Wisdom: Principles and Mechanisms" (edited by Landemore and Elster, 2012). For an exhaustive treatise on robust economic decisions of individuals who don't trust their own models, see Hansen and Sargent (2008).

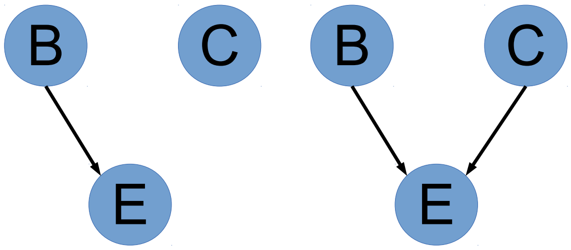
Now, I will derive the JSD from recent work in cognitive science on how humans uncover causal structure in their universe of concern (Tenenbaum & Griffiths, 2001, Griffiths & Tenenbaum, 2005) and I will show how the class of models discussed in Hansen and Sargent is related to the class of models that I use here. After that, I will show how the JSD can be generalized for multiple models and I will argue that, with the appropriate normalization, the square-root of the resulting quantity–known as the n-point JSD, or generalized JSD–is a meaningful metric of group level cognitive diversity.

2 Causal Beliefs and Joint Distributions

Causal reasoning is important for discussions in business, politics and sustainable development. Thus, for now I restrict my models of human reasoning to the causal domain, although I acknowledge that other forms of reasoning (ontological, deontic, deontological, etc.) often play roles and must be incorporated to account completely for the diversity of individual reasoning within collectives. Measures of cognitive distance and diversity are constructed not from data of cognition per se, as cognition at this scale is unobservable, but from observable reasoning or model building. The data can come from laboratory experiments in which participants build models to either answer questions or to trade on some market. Or it can come from

Cammack (2013) convincingly dispelled this interpretation of Aristotle's text and showed that Aristotle, very likely, had something very different in mind.

[4] In the case of Condorcet's jury theorem the argument is simply numerical and has little to do with people's cognition at all.



(a) Graph 0 (b) Graph 1

Figure 1: Two examples of Simple Cognitive Maps. Here "B" stands for Background Cause and "C" stands for Cause of interest (the variable which actor 1 believes has a causal effect on "E", but actor 0 does not).

interviews as well as transcriptions of speeches or debates. Transcriptions or interviews–for example–are used to elicit statements of the form "CO_2 causes Climate Change," which are then encoded as directed signed arcs of the graph representing a particular speaker's belief system as depicted in Figure 1:

$$\mathrm{CO_2} \xrightarrow{+} \text{Climate-Change.}$$

The downsides of using transcriptions of debates and speeches are that 1) it can be quite tedious to encode thousands of causal statements in the abscence of automation[5] and 2) the exact parameters of the causal structures are arbitrary. Hence, for the time being–while we are waiting for better language models–I recommend using experiments.

To begin with, people are assumed to have some objects to think about. These I call variables. For macro-economists, for example, the objects of thought would include variables such as interest and unemployement rates, GDP and inflation, among others. The researcher elicits from a set of people–perhaps macro-economists–each person's beliefs about how the variables of the domain of interest are causally related. Then, he arranges the causal relations into a coherent data structure, representing that person's belief system.

Cognitive Maps (henceforth CM), as the resulting signed digraphs are often called (see Figure 1 for two simple examples), can then be coded as *Bayesian Networks* (henceforth BN), which are parametric representations

[5] Crowdsourcing platforms such as Amazon Turk may help. See "Creating Speech and Language Data With Amazon's Mechanical Turk", Callison-Burch & Dredze, 2010.

of how a person believes that a set of variables is jointly distributed, while maintaining the causal interpretation. In experiments people might directly express their beliefs as fully parametrized Bayesian Networks, which eliminates a host of arbitrary coding decisions for the researcher. For an example of such an experiment, see Castner and Li, forthcoming. For an exhaustive treatise on causal models, see Judea Pearl's "Causality: Models, Reasoning, and Inference" (Pearl, 2000), for Probabilistic Graphical Models more generally see "Probabilistic Graphical Models: Principles and Techniques" by Koller and Friedman (2009) and for a detailed account of how beliefs are elicited and cognitive maps are constructed from people's statements see "Structure of Decision: The Cognitive Maps of Political Elites" (1976), edited by Robert Axelrod. While the exact calculations in this paper take a particular form of sub-linear causation, any functional probabilistic relationship between k variables can be specified as part of a Bayesian Network.

To show how CMs are encoded as BNs, I closely follow Griffiths, Kemp & Tenenbaum (2008). For illustrative purposes, two very simple CMs are shown in Figure 1a and Figure 1b. The universe of discourse is a set of three variables: a background cause B, a potential cause of interest, C, about the effect of which on E–the effect variable–there is a dispute. The person, let me call him 0, a representation of whose belief-system (Graph$_0$) is depicted in Figure 1a believes that only B and not C causes E, while the person, whom I call 1, with beliefs represented by Graph$_1$ in Figure 1b, believes that both B and C exert a causal influence on E. For the time being, I'm assuming all believed effects to be positive, but it is later shown how to accomodate negative effects in a principled and straight forward manner. Graph$_0$ is encoded as follows as a Bayesian Network (or simply joint distribution). The joint distribution of any k variables ($k = 3$ in this case) can be written as the product of all of its conditionals and its marginals. In the case of Graph$_0$:

$$P_0(B, C, E) = P(E|B) * P(B) * P(C)$$

Note that since B is believed to cause E, the value of E is believed to depend on the value of B and thus the term $P(E|B)$ is included. However, E is believed to be independent from C and thus $P(E|B, C)$ collapses, while in Graph$_1$ the term would have to be $P(E|B, C)$:

$$P_1(B, C, E) = P(E|B, C) * P(B) * P(C).$$

In Griffiths, Kemp & Tenenbaum (2008) all variables are binary (i.e. they can only take on values 0 and 1). But this must not be so.

To incorporate Hansen and Sargent's work, which is based on continuous variables, consider that "a decision maker's model takes the form of a linear state transition law", which they formulate as

$$y_{t+1} = Ay_t + Bu_t + C\breve{\epsilon}_{t+1}, \tag{1}$$

where $\breve{\epsilon}_t$ is the error: an i.i.d. Gaussian vector process with mean 0 and identity contemporanious covariance matrix, I. y_t is a state vector and u_t is a vector of controls. In the language of Bayes Nets, we refer to u_t as a vector of action nodes. I will return to Hansen and Sargent's work soon. There I will present Hansen and Sargent's interesting relaxation of Equation 1–the class of models that represent an actor's considered misspecifications in their theory– but for now I will continue with the exposition of Griffiths, Kemp & Tenenbaum (2008) and with the assumption that our data of people's models of the world is derived from text.

When we use text data, it is difficult to justify continuous formulations, as we know too little about a person's imagined functional relationships–we only get statements of the form A causes B. We lose no information and it is more parsimonious if we represent all variables in binary form (high or low). One could see this assumption as coming from a theory of how ordenary people who don't build quantitative models think about the world–people might coarse-grain variables as either taking on a high value (1) or a low value (0). Alternatively, the values (0, 1) could be thought of as deviations from some base-line, where 0 means a decrease and 1 denotes an increase (with an innocuous assumption that the values of the underlying variables never stay exactly the same). For a positive causal relation, when B is believed to be the cause of E for example, we have:

$$P(E = 1|B = 1) > P(E = 1|B = 0),$$

which in the case of Graph$_0$, where there is only one causal variable, can simply be parameterized as:

$$P_0(E = 1|B = b) = \pi_{0,B}b,$$

so that when b equals 0, the probability of E taking on the value 1 is believed to be 0 and when b equals 1, this is believed to cause E to take on the value 1 with probability $\pi_{0,B}$. Note that the effect parameters, such as $\pi_{0,B}$, are themselves taken to be drawn from some known distributions and while Griffiths and Tennenbaum assume uniform distributions on the interval $[0, 1]$, for this paper all calculations have been done using a few specifications of the beta distribution of which the uniform distribution is a special case–we end up with two tunable parameters. The beliefs as represented in Graph$_1$, pertaining to Es dependence on both B and C ($P_1(E|B,C)$), are slightly more difficult to parameterize. A parameterization that assumes a linear dependence on both causes ($P_1(E = 1|B = b, C = c) = \pi_{1,B}b + \pi_{1,C}c$) introduces a dependence between the two parameters–in virtue of preserving the axioms of probability–which likely has not explicitly been stated as part of the person's beliefs ($\pi_{1,B}+\pi_{1,C} < 1$). Note that with more than two causes this becomes even more problematic. Thus, Griffiths, Kemp, & Tenenbaum (2008) recommend to use Pearl's 1988 Noisy-OR parameterization:

$$P_1(E = 1|B = b, C = c) = 1 - (1 - \pi_{1,B})^b(1 - \pi_{1,C})^c. \tag{2}$$

In Equation 2, we have that if b and c are both equal to 0, the probability of E taking on the value 1 is 0. If, on the other hand, b equals 1 while c equals 0, this probability is $\pi_{1,B}$ and if b equals 0 while c equals 1 this probability is $\pi_{1,C}$. Lastly, and this is the case for which things change compared to the linear parameterization, when both b and c are equal to 1, the probability of E taking on tha value 1 is:

$$P_1(E = 1|B = 1, C = 1) = \pi_{1,B} + \pi_{1,C} - \pi_{1,B}\pi_{1,C}.$$

The reason why Equation 2 was given the name "Noisy-OR", is that in the special case where $\pi_{1,B} = \pi_{1,C} = 1$, it becomes the `OR` function, so that E takes on the value 1 whenever b equals 1, or c equals 1 or both and it takes on the value 0 otherwise. In the case of believed negative causation, supposing a Graph$_2$ which is like Graph$_1$ except that C is believed to have a negative effect on E instead of a positive one, Equation 2 becomes:

$$P_2(E = 1|B = b, C = c) = 1 - (1 - \pi_{2,B})^b(1 - \pi_{2,C})^{1-c}, \tag{3}$$

where the relationship of this probability with the value taken on by C is reversed:

$$P_2(E = 1|B = b, C = 1) = P_1(E = 1|B = b, C = 0)$$

and

$$P_2(E = 1|B = b, C = 0) = P_1(E = 1|B = b, C = 1).$$

The Noisy-OR parameterization can also be derived (as in Tenenbaum and Griffiths 2001 & 2005) from a psychological theory called "causal power", that was first suggested by Cheng (1997).

2.1 A Continuous Case.

Before I move on to derive a measure of diversity over mental models, I want to point out one more interesting class of models that is subsumed within the Bayesian Network class–that of Hansen and Sargent. The contours emerge, of a broader theory of robust decisions with increased focus on a particular constrain: the subspace that deciding collectives explore in the space of all possible models.

In general Bayesian Network formulations, it is easier to think in terms of stochastic and systematic componants (King, 1989) than in terms of errors. Thus I rewrite Hansen and Sargent's *distorted model* formulation in the following way [6]. The stochastic componant is:

$$y_{t+1} \sim N(\mu, C^2 I), \tag{5}$$

with systematic componant:

$$\mu = Ay_t + Bu_t.$$

The causal structure in this problem is the same as in one of our previous problems–it looks like Graph$_1$:

$$P(y_{t+1}, y_t, u_t) = P(y_{t+1}|y_t, u_t) * P(y_t) * P(u_t),$$

where $P(y_{t+1}|y_t, u_t)$ is the Gaussian conditional density of y_{t+1}, with linear dependencies on y_t and u_t. We have to remember that all of the variables are vectors and thus the graph has many more nodes.

To allow for feedbacks and time dependent true processes, Hansen and Sargent suggest that people might guess that their models of the world with

[6] Following King, 1989, the most general Bayesian Net with the same causal structure has a stochastic componant

$$y_{t+1} \sim F(\theta, \alpha). \tag{4}$$

F is some probability distribution with auxiliary parameter α and systematic componant θ:

$$\theta = g(y_t, u_t).$$

$g(\cdot)$ is some arbitrary deterministic function. For all statistical dependencies that we admit into our Bayesian Net, we need such a formulation to calculate all conditional pdfs. These are then to be multiplied, together with the marginal probabilities of their arguments, to arrive at the joint distribution of all the relevant variables.

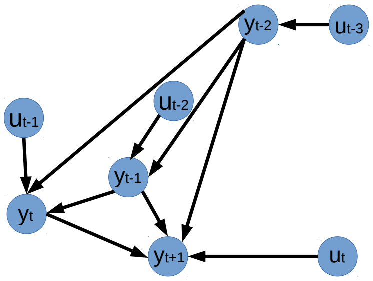


Figure 2: Here is a Graph of Hansen and Sargent's formulation with two time periods feeding back.

only limited causal possibilities (Equation 1) are misspecified and thus they surround their models with a set of alternative models of the form

$$y_{t+1} = Ay_t + Bu_t + C(\epsilon_{t+1} + \omega_{t+1}), \tag{6}$$

where ϵ_t is another i.i. Gaussian error process with mean 0 and identity covariance matrix. ω_{t+1} is a vector process that can feed back in a possibly nonlinear way on the history of y

$$\omega_{t+1} = g_t(y_t, y_{t-1}, \ldots),$$

where $\{g_t\}$ is a sequence of measurable functions. Note the generality! From their formulation, Hansen and Sargent seem to have intended to treat $\{g_t\}$ as a nuisance parameter that belongs to the stochastic componant, but it defines the difference between a theory of optimal decisions with rational expectations and Hansen and Sargent's theory of robust decisions and thus it is central. The stochastic componant of this model, then, remains the same, as in Equation 5, but the systematic componant is changed to

$$\mu = Ay_t + Bu_t + Cg_t(y_t, y_{t-1}, \ldots).$$

The structure now no longer looks like Graph$_1$–it looks like in Figure 2–and the joint distribution is as follows

$$P(y_{t+1}, y_t, y_{t-1}, \ldots, u_t) = P(y_{t+1}|y_t, \ldots, u_t)*P(y_t|y_{t-1}, \ldots)*P(y_{t-1}|y_{t-2}, \ldots)*\cdots*P(u_t).$$

Of course, y_t and u_t represent vectors of variables and each of the arrows represents a vector of influence parameters. Thus, a full causal graph taken from the class of models suggested by Hansen and Sargent can look much more complex than the one in Figure 2. The possible range of structural complexity in this class of models–the number of parameters–makes this class large. In fact, even with few variables, the number of graphs that a person has to consider for her robust decision algorithm becomes enormous. Recall that $\{g_t\}$–the part of Hansen and Sargent's models that specifies most of the causal arrows–is a sequence of arbitrary measurable functions over an infinite (or very long) sequence of vector-valued variables. Figure 2 only shows the case where the number of considered lags is 2 and even in that limited subclass note that the added number of arrows is large compared to the distorted model which looks like Graph_1. There is an important constraint that should be placed on $\{g_t\}$ and that is that causation only travels forward in time, but even with that constraint, $\{g_t\}$ is too large. Hansen and Sargent put forth that all decision makers consider Equation 1 to be a good approximation of the truth and they say that this bounds the set of models that a person has to consider. They start with an intertemporal measure of the size of model misspecification:

$$R(\omega) = 2E_0 \sum_{t=0}^{\infty} \beta^{t+1} D_{KL}(P_{g_t} || P_0),$$

where E_0 is the 0s period expectation operator of the *distorted* model, P_0 (with $\mu = Ay_t + Bu_t$) and where P_{g_t} is from the large class of models with $\mu = Ay_t + Bu_t + Cg_t$. I will define the Kullback-Leibler Divergence, D_{KL}, in the next section; suffice it to say now that it is an asymmetric measure which is often used to compare two probability distributions and that if something is not possible under P_0 that is possible under P_{g_t}, it is unbounded. $R(\omega)$ is then interpreted as a distance measure between some true process discribed by some possibly complicated causal structure indexed by ω and the distorted model with the structure of Graph_1 (Equation 1).

Hansen and Sargent's decision makers believe that the data of their experience is generated by a model that is not too far from some model in the class described by Equation 1, or Graph_1. The authors quantify this by imposing the constraint

$$R(\omega) \leq \eta_0. \tag{7}$$

Right away we see that whatever distribution has an event that is deemed impossible under P_0, that distribution is omited from consideration, even if it places only a very low probability on the offending event. But, depending on η_0, this constraint may still leave us with too many models to sift through in order to make an efficient decision in finite time. Note that during the recent housing crisis, both home owners and bankers did not believe that it was possible for housing prices to fall. Thus, their decisions were not robust–they failed to consider some events that are–as we now know–possible.

There are thus two problems: 1) Graph_1 seems arbitrary as a center around which to seach for the truth and 2) there may be many nonsensical models in this space that represent no one's beliefs. In human collectives–if we assume that people have good reasons to construct models in particular ways and that these models differ–then we have more natural constraints on the space of considered models. If each person in the collective poses a model–a point in model space–we may consider as relevant all models in the convex hull defined by these points. The resulting sub space of consideration, Ω, is likely to be smaller but also more relevant than the class considered by Hansen and Sargent's agents. The importance of this point depends on how complex reality is–how far the truth is from Equation 1. A good measure of Cognitive Diversity, measures the size of Ω. Given that people build reasonable (relevant) models, for more robust decisions in complex systems it seems that the size of Ω should be large. The more of the reasonable model space we can cover, the more reasonable concerns we capture as a collective. A reasonable model is simply one that an intelligent person might construct after careful consideration. In simple systems, like those governed by processes that are almost describable by Equation 1, there might not be many reasonable models and people's opinions might converge so that they are all situated in a tight ball around the system that Equation 1 describes exactly. In those circumstances, Hensen and Sargent's theory applies, except that too many models are considered. When the real system is complex, diversity in people's opinions is likely to persist and these are also the circumstances for which diversity is likely to make a difference. When people reason to consider models, they don't consider irrelevant ones that are not omited by Hansen and Sargent's agents, but they include relevant ones that Hansen and Sargent's agents may omit. In fact, if the truth is far from Equation 1, further than η_0, Hansen and Sargent by construction exclude it from consideration. But the restriction implied by η_0 is not based on any principle or reason. My claim is that collectives of humans can do better by employing causal reasoning. In complex systems, how well they can do likely depends on how diverse is the space of models that they can consider–it will depend on their cognitive diversity.

Another striking feature of Hansen and Sargent's models is that action nodes–or control variables–are never affected by state variables–decisions are not contingent on reality. Thus these models don't speak to theories of robust learning or adaptation and they don't allow for contingent decisions without learning or adaptation. However, we can make our actions contingent on the values of some state variables and we can anticipate to learn and adapt. In order to formulate a theory of robust learning or robust adaptation or one that allows for contingency plans we need to work with more general Bayesian Networks that allow for contingent controls.

3 Derivation of the Jensen-Shannon Divergence for Measures of Cognitive Distance and Diversity

3.1 Causal Support

Until recently (Tenenbaum & Griffiths, 2001, Griffiths & Tenenbaum, 2005) Bayesian models of human learning have typically been concerned with parameter estimation rather than with the learning of causal graph structure. However, it is the structure of people's belief systems that 1) is likely more important for understanding differences between people's beliefs and 2) is easier to obtain information about. As part of their work on causal learning–the human learning of causal structure–Tenenbaum & Griffiths have introduced a measure called *Causal Support*, which measures the support that some evidence lends to a particular structural causal theory (BN) in favor of another; it is really just a special case of a likelihood ratio, where the usual concern for parameter estimation is replaced with a concern for causal structure (or model selection):

$$\text{Support}_{1,0} = \log\left(\frac{P(D|\text{Graph}_1)}{P(D|\text{Graph}_0)}\right), \tag{8}$$

where $P(D|\text{Graph}_1)$ is the probability of seeing data D, when the data generating process is the joint distribution associated with Graph_1 and $P(D|\text{Graph}_0)$ is the probability of seeing D when the truth is described by Graph_0. This should be interpreted as the support given to Graph_1 over Graph_0 by some data D.

3.2 Cognitive Distance

Departing from Tennenbaum and Griffiths, let us now suppose that the data is repeatedly drawn from the first model specified by Graph_1 (a large number of times). The average Causal Support of that (correct) model, Graph$_1$, visá-vis another model, Graph$_0$, can then be seen as the degree to which the first model can be distinguished from the second one, if the first one in fact specifies the correct data generation process. The resulting quantity is what Hansen and Sargent use in their treatise on robust dynamic decisions and it is known as the Kullback-Leibler divergence of Information Theory (also Information Divergence, Information Gain, Relative Entropy, or KLIC):

$$D_{KL}(P_1||P_0) = E_1(\log\left(\frac{P_1(D)}{P_0(D)}\right)),\ \text{with } D \sim P_1,$$

where $E_1(\cdot)$ is the expectation operator under Graph$_1$ (not the effect variable)[7]. However, in this example as in many others, it is clear that $D_{KL}(P_1||P_0)$ is not defined, because Graph$_0$ puts zero probability on $D = (B = 0, C = 1, E = 1)$, which will in expectation be drawn $P_1(C = 1) * \pi_{1,C} * N$ times for every N draws.

The Kullback-Leibler Divergence is relevant when an assymetric measure of distance is sought, as in Baldi and Itti (2009) who define surprise as an assymetric distance between a prior and a posterior model[8].

Definition Define the amount of Surprise one bit of data D has on a Bayesian Network, G as

$$D_{KL}(G, G|D) = \int_{\chi} P(G) \log \frac{P(G)}{P(G|D)} dG,$$

where $D_{KL}(\cdot)$ is the relative entropy or Kullback-Lieber Divergence. It is a quantification of the effect that one bit of data, D had on the model G.

But unlike surprise, distance between two people in cognitive space should be symmetric and finite. While surprise is likely to be finite, if it is not it has a specific interpretation–something thought to be impossible has proven to be possible. In the case of distances between models, there

[7] For notational convenience, I write $P_1(D)$, $P_0(D)$, instead of $P_1(D|\text{Graph}_1)$ and $P_0(D|\text{Graph}_0)$.

[8] Baldi and Itti (2009) argue that for a measure of surprise, asymmetry better matches intuitive notions: "A broad prior distribution followed by a narrow posterior distribution corresponds to a reduction in uncertainty, while a narrow prior distribution followed by a broad posterior distribution corresponds to an increase in uncertainty, and both lead to different subjective experiences," where broad and narrow joint distributions are defined as higher and lower entropy distributions respectively. Entropy is the most general measure of uncertainty–for arbitrary distributions–which is guaranteed to satisfy a set of desirable axioms (Cover Thomas, 2006).

could be many things that one person holds as possible and another person holds as impossible. If we end up with an infinite distance every time one person finds something impossible that another finds possible, then we can not distinguish between pairs of people who have multiple such disagreements and those who only disagree about the possibility of one event. Further, the measure of distance must be less sensitive to zero probability beliefs than the Kullback-Lieber Divergence, if it is needed to distinguish between pairs of people who completely disagree about the possibility of an event–one finds it likely, while the other finds it impossible–and pairs of people who almost agree in that one person finds the event again impossible while the other finds it *almost* impossible.

To guarantee that our distance measure is finite and symmetric, let $M = \lambda P_1 + (1 - \lambda)P_0$ denote the mixture of the two joint distributions, with $\lambda \in (0, 1)$. It is then guaranteed that $D_{KL}(P_1||M)$, the average causal support of Graph$_1$, vis-á-vis the mixture, M, when Graph$_1$ generates the data, takes on finite values–for all the calculations in this paper $\lambda = \frac{1}{2}$ because it makes the JSD symmetric. The same can then be done in reverse where the average causal support of the second model, Graph$_0$, over the mixture M, is calculated with data repeatedly drawn from the distribution specified by Graph$_0$:

$$D_{KL}(P_0||M) = E_0(\log\left(\frac{P_0(D)}{M(D)}\right)), \text{ with } D \sim P_0.$$

The JSD for the two models is then obtained by taking a weighted average over these two expectations:

$$\text{JSD}_\lambda(P_1||P_0) = \lambda D_{KL}(P_1||M) + (1-\lambda)D_{KL}(P_0||M). \tag{9}$$

This quantity, JSD$_\lambda$, can be interpreted as the average distinguishability between two joint distributions (cognitive models in this case) given one bit of data (DeDeo, Hawkins, Klingenstein and Hitchcock 2013)–it measures the total divergence to the average or the Information Radius (IRad). The information radius quantifies the amount of information that is lost if we were to describe two processes, P_1 and P_2, by their average M. Also note that this measure, unlike D_{KL}, is symmetric when λ is set to $\frac{1}{2}$ (I drop the λ subscript when $\lambda = \frac{1}{2}$):

$$\text{JSD}(P||Q) = \text{JSD}(Q||P), \forall P, Q.$$

When in addition the base 2 logarithm is used we have that

$$0 \leq \text{JSD}(P||Q) \leq 1, \forall P, Q$$

.

Lastly, if we take the square root of this quantity, we obtain a metric so that the triangle inequality holds (Endres and Schindelin, 2003). One implication is that with this metric the Banach fixed-point theorem holds for the model space (Palais, 2007). Thus we have a symmetric and finite measure that is defined everywhere, satisfies the triangle inequality and is non-negative, vanishing only when $P = Q$–these are all important properties, if the measure is meant to compare distances between different belief systems so that we can make statements such as “the distance between Q and P is larger than that between P and R”. The resulting metric can then be interpreted as the “cognitive distance” (CD) between any two models:

$$\text{CD}(P||Q) = \sqrt{\text{JSD}(P||Q)}. \tag{10}$$

3.3 Cognitive Diversity

An extension to the JSD–the n-point Jensen-Shannon Divergence, or JSD$_n$–can be defined as:

$$\text{JSD}_n(P_1, P_2, \ldots, P_n) = \sum_{i=1}^{n} \omega_i D_{KL}(P_i|\bar{P}), \tag{11}$$

where $\bar{P}$ is the mixture of all n models and where the ω_is are weights that sum to 1. In the special case where all distributions are defined over a finite set, Equation 11 can be expressed as

$$\text{JSD}_n(P_1, P_2, \ldots, P_n) = H\left(\sum_{i=1}^{n} \omega_i P_i\right) - \sum_{i=1}^{n} \omega_i H(P_i), \tag{12}$$

where $H(\cdot)$ is defined as the Shannon Entropy. In the case that $\omega_i = \frac{1}{n}$, $\forall\ i$, Gallager (1968) proved that the JSD$_n$ is a convex function in $(P_1, P_2, \ldots, P_n)$. This measure can be interpreted as the amount of information that is gained from one arbitrary data sample, about which among the n distributions is the closest one to the underlying true distribution describing the system. Note that from the outset, a measure was sought that would measure a collective’s cognitive diversity as the size of its current available theoretical tool kit and the JSD$_n$ has this quality; theoretical distributions are compared to a data sample and the greater is the numerical value of the JSD$_n$, the more theoretical material there is among the models to make sense of the data.

3.3.1 A Correction for Group Size and Redundancy

Note that the maximum value of the JSD_n–which I call the potential diversity–is an increasing and concave function of n as should be intuitive, or more precisely it is:

$$\text{Potential Diversity}(n) = \log_2(n), \tag{13}$$

which in the special case where $n = 2$ is equal to 1. It follows from Equation 11, that $0 \leq \text{JSD}_n \leq \log_2(\text{n})$.

Further, for any tuple of probability distributions–for example the two-tuple $(P,\ Q)$–the JSD_n has the same numerical value over just that tuple as the JSD_{nk}, over any tuple with multiples of this tuple's entries (k times the same entries as in the original tuple, where k is a positive integer). For example $\text{JSD}_2(P, Q) = \text{JSD}_4(P, P, Q, Q)$. Taken together, Equation 13 and this last point represent a problem for a measure of diversity! Two different view-points in a collective of two people seems to mean, intuitively, that the collective of two is diverse, while having two even radically different view-points in a collective of a hundred people makes the collective decidedly less diverse. The measure–so far–violates this intuition. But luckily, there is an easy fix: correct every comparison for group size by deviding the JSD_n by its maximum potential value, $\log_2(n)$:

$$\text{Cognitive Diversity}_n(P_1, P_2, \ldots, P_n) = \sqrt{\frac{1}{\log_2(n)}\text{JSD}_n(P_1, P_2, \ldots, P_n)}. \tag{14}$$

The cognitive diversity as defined in Equation 14, as it is normalized for group size, allows to compare the cognitive diversity of collectives with different numbers of individuals, as well as collectives with an equal number of individuals. With this measure, as it discounts the diversity of a group increasingly with group size, a greater number of people is not likely to lead to a greater magnitude in diversity and thus, unlike most existing instruments of diversity research, it is not meant as a tool with which to ask questions about the absolute diversity of any group. It is meant to ask questions related to a group's diversity, relative to how diverse it could be. We can then seperate group size and diversity and if we include both in our analysis, we can see how each seperately affects the quality of a collective's decisions–provided that we have a measure for decision quality.

4 Causal Explanations of the 2008 Foreclosure Crisis

Since the economic crisis[9] commenced in 2008, many narratives seeking to explain the onset of this costly phenomenon have appeared in public discussions (including four congressional committees), speeches, newspaper articles, academic papers and books. I thus use the statements of a few important analysts of the crisis to show how beliefs can be represented as Bayesian Networks and how the cognitive diversity of a collective of such experts is then approximately measured[10]. The material, alongside some explanations of how it is used to construct the cognitive maps, exhibited in Figure 3, can be found in the Appendix, but I give an example here: In the Financial Crisis Inquiry Commission Staff audiotape of the interview with Warren Buffett on the 26th of May, 2010, Mr. Buffet was recorded as saying the following:

> The basic cause was, you know, embedded in, partly in psychology, partly in reality in a growing and finally pervasive belief that house prices couldn't go down. And everybody succumbed, virtually everybody succumbed to that. But that's, the only way you get a bubble is when basically a very high percentage of the population buys into some originally sound premise ...

As Mr. Buffett only spoke of one cause, I will humerously name it CD: Cognitive Diversity, (virtually everybody succumbed) and as Mr. Buffet blamed the onset of the crisis on the lack of CD, I arrive at Mr. Buffet's CM (Figure 3h).

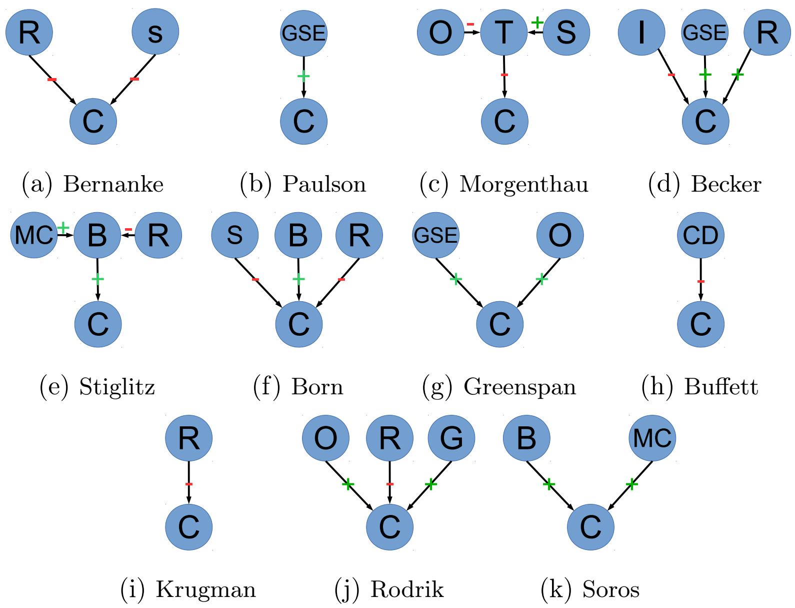
4.1 The Cast of Characters

- Ben Bernanke (economist and chairman of the US Federal Reserve Bank, Figure 3a),

- Henry Paulson, Jr. (past CEO of Goldman Sachs and Secretary of the US Treasury at the time of the crisis, Figure 3b),

[9] As it is generally accepted that the economic crisis was the result of a housing foreclosure crisis, or subprime mortgage crisis, these terms are here used as if they were interchangable. This is a simplification that should not matter, as one could add to every belief system the same extension, which has the same effect on the discussed measures as if this extension is simply collapsed into one effect node which includes all of these terms.

[10] The python code for this exercise can be found on my github site.



(a) Bernanke (b) Paulson (c) Morgenthau (d) Becker

(e) Stiglitz (f) Born (g) Greenspan (h) Buffett

(i) Krugman (j) Rodrik (k) Soros

Figure 3: C: Crisis, R: Regulation, S: Supervision, I: Interest Rate, T: Transparency, O: Offshoring, GSE: Government Promotion of Home-Ownership, B: Banking Behavior, MC: Misguided Incentives, CD: Cognitive Diversity

- Robert Morgenthau, (District Attorney for New York County at the time of the crisis, Figure 3c),
- Joseph Stiglitz (economist, Figure 3e),
- Brooksley Born (Commissioner on the Financial Crisis Inquiry Commission and past chair of the Commodity Futures Trading Commission, Figure 3f),
- Alan Greenspan (economist and past chairman of the US Federal Reserve Bank, Figure 3g),
- Warren Buffett (CEO and largest shareholder of Berkshire Hathaway, known as the “Oracle of Omaha”, Figure 3h),
- Paul Krugman (economist, Figure 3i),
- Dani Rodrik (economist, Figure 3j)

and

- George Soros (Chairman of Soros Fund Management and philanthropist, Figure 3k).

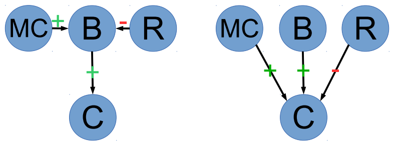
4.2 Cognitive Distances

Committee	Bernanke	Paulson	Morgenthau	Becker	Stiglitz	Born	Greenspan	Buffett	Krugman	Rodrik	Soros
Bernanke	0.0										
Paulson	0.3483	0.0									
Morgenthau	0.3791	0.3778	0.0								
Becker	0.3592	0.3394	0.4345	0.0							
Stiglitz	0.3965	0.3783	0.4295	0.4651	0.0						
Born	0.1823	0.3968	0.4209	0.3162	0.3999	0.0					
Greenspan	0.3104	0.2415	0.3791	0.2636	0.4134	0.3151	0.0				
Buffett	0.3483	0.3128	0.3778	0.3968	0.3783	0.3968	0.3483	0.0			
Krugman	0.2415	0.3128	0.3778	0.447	0.3465	0.3394	0.3483	0.3128	0.0		
Rodrik	0.2636	0.3968	0.4209	0.3162	0.4472	0.238	0.2636	0.3968	0.3394	0.0	
Soros	0.3104	0.3483	0.399	0.3151	0.3198	0.2636	0.3104	0.3483	0.3483	0.3151	0.0
After adjustment for Joseph Stiglitz:											
Stiglitz	0.2636	0.3968	0.4345	0.3162	0.0	0.1764	0.3151	0.3968	0.3394	0.238	0.1823

Table 1: The pair-wise cognitive distance measure, $\sqrt{\text{JSD}_2(i,j)}$, for each pair of experts.

Table 1 shows the Cognitive Distance, $\sqrt{\mathrm{JSD}_2(i,j)}$, between any two experts. The five largest distances are, in order from greatest to lowest magnitude, those between 1) Stiglitz and Becker (0.465), 2) Stiglitz and Rodrik (0.447), 3) Krugman and Becker (0.447), 4) Morgenthau and Becker (0.4345) and 5) Stiglitz and Morgenthau (0.429). The cognitive maps of both, Joseph Stiglitz (Figure 3e) and Gary Becker (Figure 3d) are present in three out of the five largest diadic distances, while that of Robert Morgenthau (Figure 3c) is involved in two of the five largest distances. The maps of Stiglitz and Morgenthau are structurally more complex than all others in the collection, in that they both include a mediating variable through which two other variables causally affect the onset of the crisis, instead of being composed of 1 to 3 direct causes which is the case of all other maps. Also, Morgenthau's map includes a variable, T (transparency) that is absent from all other maps, while Stiglitz's map includes a variable MC (misguided incentives) which is present in only one other map (Soros's, Figure 3k). Gary Becker's map is unique in that it includes a *positive* causal relation from R (financial regulation) to the onset of the crisis, C, while all others who considered R argued that its lack was responsible and not that there was too much of it. Becker stated that the regulators were in part to be blamed for the crisis, as they were "cheerleaders for the banks," and it is important to note that my choice to code Becker's partial blame on the regulators as a positive causal relation from R to C is debatable. Indeed, R might not be the right variable, if R is the symbol that is used for all others to denote the quantity of regulation, and what might be needed is an additional variable RB (the behavior of the regulators). In order to keep a bound on the number of variables (to keep things simple) I choose to code Becker's statement as $R \xrightarrow{+} C$, with the explicit caveat that this assumption might cause to exaggerate the magnitudes of some of my measures.

The five shortest distances, in order of increasing magnitude, are between 1) Bernanke and Born (0.182), 2) Rodrik and Born (0.238), 3) Bernanke and Krugman (0.241), 4) Greenspan and Paulson (0.241) and 5) Becker and Greenspan, Rodrik and Greenspan, Bernanke and Rodrik and Born and Soros (all with distance 0.2636). The shortest cognitive distance is that between Ben Bernanke (Figure 3a) and Brooksley Born (Figure 3f), whose cognitive maps are essentially the same, except that Born's map includes one additional positive edge from B, the behavior of the banks, to the onset of the crisis, C. The second shortest cognitive distance, that between Dani Rodrik (Figure 3j) and Brooksley Born (Figure 3f), is already much greater in magnitude; it is by a factor of 1.3 greater than the smallest, where the



(a) Stiglitz (before adjustment)

(b) Stiglitz (after adjustment)

Figure 4: A plausible simplification of Joseph Stiglitz's CM, variables are: C: Crisis, R: Regulation, B: Banking Behavior, MC: Misguided Incentives.

maximum distance is by a factor of about 2.6 greater. This jump in magnitude, from the shortest to the second shortest distance is no surprise when one looks at the two graphs involved in the calculation; Rodrik's and Born's maps have one causal relation in common and are similar in structure, but each has two causes that the other has not.

4.3 Sensitivity of the Measures

It is important to experiment with these measures in order to get a better understanding of their meaning. For example, with these representations of beliefs, Joseph Stiglitz might be further removed in distance than is truly warranted, from Born, Bernanke, Rodrik and Krugman, simply because his map includes behavior as a mediating variable, mediating between incentives as well as regulation and the onset of the crisis, where the others very likely have the same in mind but see this as too trivial to make explicit (hence their maps look very different). Making an adjustment that simplifies Joseph Stiglitz's map (see Figure 4), decreases the overall diversity measure, from 0.302 to 0.289. For Joseph Stiglitz, his distance to Bernanke decreases to 0.264, his distance to Paulson increases to 0.397, while the decrease of his distance to Born is most dramatic, decreasing from 0.4 to 0.18, an adjustment which makes them the closest in terms of cognitive distance for the whole collection. Thus, it is clear that these measures are very sensitive to the exact specification of beliefs and that therefore great care must be taken in the elicitation and processing of people's statements. However, I see this sensitivity as a strength, rather than a weakness of the measuring approach, as the diversity that results from differences in exact communication patterns and thoughts (such as the inclusion and exclusion of potentially important mediating variables), might be precisely what leads to a collective's greater understanding of the world.

4.4 Constructing Diverse Collectives

Committee	Bernanke	Paulson	Morgenthau	Becker	Stiglitz	Born	Greenspan	Buffett	Krugman	Rodrik	Soros
$n = 11$	-0.0112	-0.0055	0.0054	-0.0018	0.0058	-0.0091	-0.0107	-0.003	-0.006	-0.007	-0.0099
$n = 10$		-0.0078	0.0049	-0.0036	0.0049	-0.0094	-0.0135	-0.0047	-0.0067	-0.0081	-0.0125
$n = 9$		-0.009	0.004	-0.0037	0.0027	-0.0125		-0.0073	-0.0099	-0.0094	-0.0161
$n = 8$		-0.0133	0.0015	-0.0049	0.0031	-0.015		-0.0111	-0.0145	-0.0126	
$n = 7$		-0.0206	-0.0016	-0.0037	0.002			-0.0175	-0.0192	-0.0116	
$n = 6$			-0.0053	-0.0041	-0.0002			-0.0236	-0.026	-0.0213	

Table 2: This table represents the algorithm of iterated deletion of diversity minimizing elements (the algorithm is as in Equation 15). Morgenthau, Becker, Stiglitz, Buffett, Krugman and Rodrik survived the iterated deletion of diversity minimizing elements, for a maximally diverse group of 6.

Interesting is also to measure how much each individual view of the crisis contributes to the diversity of the collection of views, so that an l person team of experts can be constructed with the goal of maximizing diversity in mind (if that were to be found desirable)[11]. There are two ways in which a maximally diverse group of, say 6, could be constructed from a group of 11: one way is to repeatedly subtract that person from the group whose presence contributes the least to (or subtracts the most from) the diversity of the group, (i.e. Equation 15) and the other is to, starting from the cognitive distance of two people's graphs, repeatedly adding that additional person whose inclusion maximizes the cognitive diversity of the larger group (Equation 16):

$$\min_i \left(\sqrt{\frac{JSD_n(\Omega_n)}{\log_2(n)}} - \sqrt{\frac{JSD_{n-1}(\Omega_n \setminus \text{Graph}_i)}{\log_2(n-1)}} \right), \text{ for } n = 10, \ldots, 6, \quad (15)$$

where Ω is the collection of all graphs and $\Omega \setminus \text{Graph}_i$ is the collection of all graphs, except Graph_i: the graph whose exclusion maximizes the diversity over the remaining $n-1$ graphs (see Table 2 for an illustration).

$$\max_i \left(\sqrt{\frac{JSD_{\tau+1}(S_\tau \oplus \text{Graph}_i)}{\log_2(\tau+1)}} - \sqrt{\frac{JSD_\tau(S_\tau)}{\log_2(\tau)}} \right), \text{ for } , \tau = 1, \ldots, 4, \quad (16)$$

[11] In practice of course, there are many more conciderations aside from just cognitive diversity and it is likely never advisable to be entirely directed by such a uni-dimensional goal.

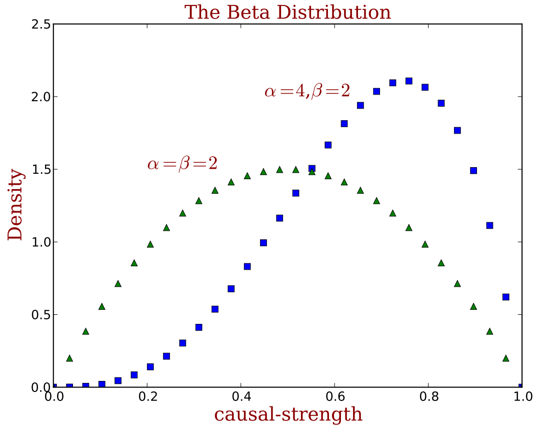


Figure 5: The beta distribution for two different parameterizations (α, β).

where S is initiated as one of those elements, with the largest cognitive distance in the group to some other element (here Joseph Stiglitz, before the adjustment) and then is incremented each time to maximize the diversity of $S_\tau \oplus \text{Graph}_i$, the collection that includes all members in S and the additional member i, whose graph maximizes the diversity of the resulting collection with size $\tau + 1$.

4.5 Causal Intensity: the Parameters

Recall that for any belief system (Graph_i), the probability of a data point, D, given the beliefs is calculated as

$$P(D|\text{Graph}_i) = \int_0^1 \dots \int_0^1 P_i(D|\text{Graph}_i, \pi_{i,0}, \dots, \pi_{i,k}) P(\pi_{i,0}, \dots, \pi_{i,k}|\text{Graph}_1) d\pi_{i,0} \dots d\pi_{i,k}$$

for k causal effect parameters, where the “Noisy-OR” parameterization is used. The effect parameters themselves are drawn from the joint-distribution, $P(\pi_{i,0}, \dots, \pi_{i,k}|\text{Graph}_i)$, which in this case is simply the product of the marginals (I assume parameters to be drawn independently from their marginal distributions). Further, as a speaker's emphasis is harder to evaluate, I assume all effect parameters to be drawn from the same beta distribution, $B(\alpha, \beta)$ with shape parameters, α and β (see Figure 5). The greater both parameters are in value, the smaller is the variance of the beta distribution and the greater is the ratio, $\frac{\alpha}{\beta}$, the greater is the density for believed causal effects closer to 1. These parameters, of course, also effect the magnitude of the diversity measure and its sensitivity. Before the adjustment of Stiglitz's belief system, the diversity increases from 0.316 to 0.45 if α is changed from 2 to 4 while β is held constant and after the Stiglitz adjustment, it changes from 0.289 to 0.403. Since the difference between 0.45 and 0.403 is comparable in magnitude (judged by the relative magnitudes of the pairwise distances) to the difference between 0.316 and 0.289, α does not seem important in ordinal terms (i.e. if the goal is to judge between group differences in diversity). If the goal is to judge the diversity between structural beliefs as accurately as possible (having only information about structure and not about believed causal strength), it is advisable to choose higher αs and βs, as well as higher ratios, $\frac{\alpha}{\beta}$, as that makes the measures more sensitive to smaller structural differences (it also assumes people to be more certain and to have stronger beliefs). Of course, if more information is available about the strengths of individual beliefs, α and β can be adjusted so as to take this information into account.

5 Conclusion

By connecting ideas from various disciplines; cognitive science (Griffiths and Tenenbaum, 2001, 2003, 2005, 2008), political Science (Axelrod 1976) and information theory (DeDeo et. al 2013), this paper demonstrates how a theory of human causal learning, naturally gives rise to some meaningful measures. I show how these measures may be combined with texts from utterances of a collective's members, to measure that collective's cognitive diversity. Using recent opinion pieces and testimonies about the 2008 financial crisis as an example data set, I describe and demonstrate "hiring and firing" algorithms, if cognitive diversity were to be seen as a goal. I see this paper as a first step toward a theory of robust collective decisions that can be confronted with empirical data.

6 References

Anderson John R. 2008. *Cognitive Psychology and its Implications.* Worth Publishers; Seventh Edition edition.

Axelrod R. (ed) (1976). *Structure of decision : the cognitive maps of political elites.* Princeton : Princeton University Press.

Converse PE. 1965. The Nature of belief systems in mass publics, In *Ideology and discontent*, ed. Apter, D.E. New York: Free Press.

DeDeo S., R. Hawkins, S. Klingenstein, and T. Hitchcock (2013). *Bootstrap methods for the empirical study of decision-making and information ows in social systems.* eprint arXiv:1302.0907, December 2013. http://arxiv.org/abs/1302.0907][http://arxiv.org/abs/1302.0907http://arxiv.org/abs/1302.0907]]. Entropy, in press.

Gallager R. G. (1968) “Information Theory and Reliable Communication,” Wiley, New York.

Griffiths T. L., Kemp, C., and Tenenbaum, J. B. (2008). *Bayesian models of cognition.* In Ron Sun (ed.), Cambridge Handbook of Computational Cognitive Modeling. Cambridge University Press.

Griffiths T.L., & Tenenbaum, J.B. (2005). *Structure and strength in causal induction.* Cognitive Psychology 51, 334-384. (This paper was formerly titled “Elemental causal induction.”)

Hong L., Page S. (2004) *Groups of diverse problem solvers can outperform groups of high-ability problem solvers.* Proceedings of the National Academy of Sciences 101(46): 16385–16389.

Koller, D. and N. Friedman (2009). Probabilistic Graphical Models: Principles and Techniques. edited by . MIT Press.

Landemore H., Elster J. (eds) (2012). *Collective wisdom: Principles and mechanisms.* Cambridge University Press, Cambridge.
Lombrozo T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, Vol. 10(10): 464-470.

Palais, R. "A simple proof of the Banach contraction principle." J. fixed point theory appl. 2 (2007), 221–223.

Pearl, J. (1988)., *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA.

Tenenbaum J. B., T. L. Griffiths (2003), *Theory-based causal inference.* Advances in Neural Information Processing Systems 15. Becker, S., Thrun, S., and Obermayer. (eds). Cambridge, MIT Press, 2003, 35-42.

Tenenbaum J. B., T. L. Griffiths (2001), *Structure learning in human causal induction.* Advances in Neural Information Processing Systems 13. Leen, T., Dietterich, T., and Tresp, V., Cambridge, MIT Press, 2001, 59-65.

Zaller J. (1991). Information, Values and Opinion. *The American Political Science Review*, Vol 85(4):1215-1237.