Measures of Cognitive Distance and Diversity.

Johannes Castner

November 8, 2013

Contents

Abstract

This paper explains and demonstrates how a model of human causal learning, *Causal Support* (Tenenbaum & Griffiths, 2001, Griffiths & Tenenbaum, 2005), can be used to derive an Information theoretic quantity, the Jensen-Shannon Divergence, with which to measure *Cognitive Distance*, the degree to which two people differ in their opinions about the workings of the world. It is then shown how this measure is generalized to measure *Cognitive Diversity*, the degree of heterogeneity of opinions within a collection of individuals, such as a political party or a research department of a firm or university. These measures are important for theoretical and empirical work on the relationships between Cognitive Diversity and an organization's success in recognizing structure in the universe of interest and in making collective decisions.

1 Introduction

An understanding of the conditions for collective wisdom and the mechanisms that enable collectives to become wiser, are at the core of what we might hope to call a science of sustainable development. Most decisions that are relevant to sustainable development, however we may define it, are at least indirectly influenced by human collectives (i.e. committees, congress, corporate boards etc.). One important set of machanisms relates the diversity of individual minds within a collection to the wisdom with which it solves its problems (collective decision and collective action). In economics and political science much has been done to uncover important mechanisms with respect to heterogenous preferences, information asymmetries and even the diversity of received signals (information) when beliefs are homogenous, in market settings and strategic games. However, the situation in which information and goals are the same among members of a collection but beliefs are heterogenous (such as in a team), has to my knowledge not been given much attention and neither have those situations, which are likely to be quite common in our world, where both beliefs and goals, or both beliefs and information-selection are diverse (people endogenously select their sources). Climate change discussions at all scales (from local to global) for example, are likely to draw attendees with highly diverse goals and beliefs, who also are exposed to very different information.

Starting with Condorcet's jury theorem, (Marquis de Condorcet, 1785)[1], 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) and it is this diversity of minds and not the sheer number of people that is believed to lead to better collective decisions[2]. Analogous with the definition of bio-diversity as the amount of genetic information that is available to the system (for greater functional diversity under a greater range

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

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

of circumstances), an organization's cognitive diversity can be defined as the collective's current available theoretical tool kit and it should thus increase the degree of accuracy with which the collective will judge the workings of the world, under a greater number of contingencies[3]. This, in turn, should improve the collective interaction of the organization with the world; particularly if the world of interest is constantly changing. Models of the world (to be made precise below) that had been inaccurate previously, might have more explanatory power as circumstances change, while those models that once furnished the best explanations might fade in relevance. For a thorough discussion and overview of the literatures on and related to collective wisdom, distributed intelligence/cognition or the wisdom of crowds, see "Collective Wisdom: Principles and Mechanisms" (edited by Landemore and Elster, 2012).

In much, but not all of this and related work, the number of people in a group of thinkers acts as a proxi for diversity in the absence of a direct measure, which leaves open all questions concerning the differential cognitive diversity in groups with more or less the same fixed number of members. Although diversification of investment in resources and capital has long been standard advise in portfolio theory, not much is known or recommended in terms of the diversification of investment in human resources, with the notable exception of some recent theoretical work by Hong and Page (2004) which has inspired the present work. Additionally, this research direction could give rise to some nuanced and unorthodox policy recommendations for the educational sector. For example, it may result in the advise to expose different students (at all ages and levels) to varying educational opportunities, suited to their own individual needs and desires.

Socially constructed measures of diversity (along ethnic, gender and religious dimensions, for example), the definitions of which are highly sensitive to context and interpretation, have become a common public agenda of firms and organizations. This diversity agenda is promoted mostly 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. I content that the current lack of a natural measure of cognitive diversity stands in the way of constructing falsifiable theories directly relating the cognitive diversity of a collective to its collective wisdom and that this lack leads to an impasse which helps

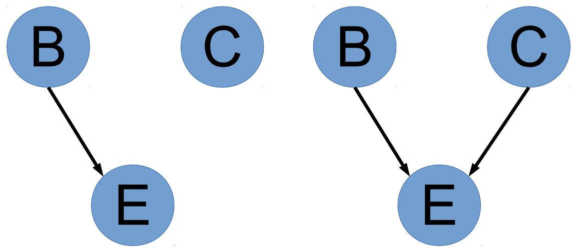
[3]Whether and to what degree this is true for any given organization must depend also on the organization's opinion aggregation scheme, just as the benefits of bio-diversity depend on the structure of the food web.

to explain the current dirth in the theory and empirics of this important subject. 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. Thus, a focus on these often more easily apparent forms of diversity might be justifiable not only on ethical or esthetic grounds, but also on grounds of organizational efficacy and robustness, if cognitive diversity can indeed be scientifically shown to have such benefits. As a requisite for such scientific work, however, adequate and theoretically well-motivated measures of *cognitive distance* (between any two individuals) and *cognitive diversity* (for a collection of individuals) are needed[4]. Thus, this paper concerns itself with such measures. The here advanced cognitive distance metric is the square root of a measure known as the Jensen-Shannon Divergence (henceforth JSD). JSD is an important information theoretic quantity that will here be derived from recent work in cognitive science on how humans uncover causal structure in their universe of concern (Tenenbaum & Griffiths, 2001, Griffiths & Tenenbaum, 2005). JSD is then generalized and the square-root of the resulting quantity (known as the n-point JSD, or generalized JSD), with the apropriate normalization, is argued to be a meaningful metric of group level cognitive diversity.

2 Causal Beliefs and Joint Distributions

Underlying many of the important discussions in business, politics and sustainable development are matters of causality and causal reasoning. Thus, as a preliminary step, this paper restricts itself to the causal domain, although it is acknowledged that other forms of reasoning (ontological, deontic, deontological, etc.) quite often play important roles and must be incorporated to account completely for the diversity of individual reasoning within a collective. Here, "reasoning" replaces "cognition" to temporarily draw attention to the fact that the 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. Either interviews or transcriptions of speeches/debates are used to elicit statements of the form "CO_2 causes Climate Change," which are then encoded as directed signed arcs of the graph

[4] If the mechanisms by which cognitive diversity impacts performance and robustness can then be sufficiently isolated and it can be shown that this form of diversity indeed correlates with the other forms, it may be found simpler and more convenient to use more easily visable proxies in actual practice (for example, in human resource departments of firms).



(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).

representing a particular speaker's belief system as depicted in Figure 1:

$$\mathrm{CO}_2 \xrightarrow{+} \text{Climate-Change}.$$

For macro-economists, for example, the universe of discourse would include variables such as interest rates, unemployement, GDP and inflation, among others. *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 structural representations of how a person believes that a set of variables is jointly distributed, while maintaining the causal interpretation. 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 are restricted to arguments of "linear causation"[5], i.e. unconditional causal effects, making causal effects conditional on the values of various variables is a straight-forward generalization that is exhaustive of all possibilities, as far as causation is concerned[6]. There are two possibly related reasons why linear causal beliefs are of special interest; 1) in human discussions (the data), people tend to express themselves in terms of positive and negative causal relations (Axelrod et. al 1976) and it would be absurd to introduce arbitrary functional relations between

[5]The exact functional relationship as specified by the intuitive Noisy-OR parameterization is sub-linear for multiple causes, to be more precise.

[6]Any functional probabilistic relationship between any k variables can be specified as part of a Bayesian Network, but such an excercise is not the point of this paper.

variables as part of characterizing people's beliefs and 2) recent theories and empirical results in cognitive science suggest that people are indeed simple in their beliefs of how variables affect each other, which, as shown in Griffiths and Tennenbaum (2005), naturally leads to Pearl's "Noisy-OR" parameterization (Pearl, 1988), described in detail below.

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 $P(E|B, C)$:

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

I assume that all variables are binary (i.e. they can only take on values 0 and 1). One could see this assumption as coming from a theory of how people think, where people are theorized to coarse-grain variables as either taking on a high value (1) or a low value (0), or 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 beta distributions as they are more flexible. The beliefs as represented in Graph$_1$, pertaining to E s dependence on both B and C ($P_1(E|B, C)$), are slightly more difficult to parameterize. A parameterization that assumes a simple linear d ependence on both causes ($P_1(E = 1|B = b, C = c) = \pi_{1,B}b + \pi_{1,C}c$) would introduce a dependence between the two parameters which likely has not explicitly been stated as part of the person's beliefs ($\pi_{1,B} + \pi_{1,C} < 1$), in virtue of preserving the axioms of probability. Note that with more than two causes this becomes even more problematic. Thus, I follow the recommendation in Griffiths, Kemp, & Tenenbaum (2008) in using Pearl's 1988 Noisy-OR parameterization:

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

In Equation 1, we have that if b and c are both equal to 0, the probability of E taking on the value 1 is 0, if only 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 1 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 1 becomes:

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

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).

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 causal induction 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 to understanding differences between people 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 of parameter estimation is replaced with a concern for causal structure (or model selection):

$$\text{Support}_{1,0}=\log\left(\frac{P(D|\text{Graph}_1)}{P(D|\text{Graph}_0)}\right), \tag{3}$$

which should be interpreted as the support given to Graph_1 over Graph_0 by some data D. $P(D|Graph_1)$ and $P(D|Graph_0)$ are computed by integrating over the parameters associated with the two different structures:

$$P(D|\text{Graph}_1)=\int_0^1\int_0^1 P_1(D|\text{Graph}_1,\pi_{1,B},\pi_{1,C})P(\pi_{1,B},\pi_{1,C}|\text{Graph}_1)d\pi_{1,B}d\pi_{1,C}$$

and

$$P(D|\mathrm{Graph}_0) = \int_0^1 P_0(D|\mathrm{Graph}_0, \pi_{0,B}) P(\pi_{0,B}|\mathrm{Graph}_0) d\pi_{0,B},$$

where $P(\pi_{1,B}, \pi_{1,C}|\mathrm{Graph}_1)$ and $P(\pi_{0,B}|\mathrm{Graph}_0)$ are the distributions from which the parameters are believed to be drawn (uniform, or beta distributions, for example). We then obtain:

$$\mathrm{Support}_{1,2} = \log\left(\frac{\int_0^1\int_0^1 P_1(D|\mathrm{Graph}_1, \pi_{1,B}, \pi_{1,C}) P(\pi_{1,B}, \pi_{1,C}|\mathrm{Graph}_1) d\pi_{1,B} d\pi_{1,C}}{\int_0^1 P_0(D|\mathrm{Graph}_0, \pi_{0,B}) P(\pi_{0,B}|\mathrm{Graph}_0) d\pi_{0,B}}\right).$$

3.2 Cognitive Distance

Departing from Tennenbaum and Griffiths now, 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, vis-á-vis another model, Graph_0, can then be seen as the degree to which the first model can be destinguished from the second one, if the first one in fact specifies the correct data generation process. The resulting quantity is known as the Kullback-Leibler divergence (also Information Divergence, Information Gain, Relative Entropy, or KLIC):

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

where $E_1(\cdot)$ is the expectation operator under Graph_1 (not the effect variable). For notational convenience, I write $P_1(D)$, $P_0(D)$, instead of $P_1(D|\mathrm{Graph}_1)$ and $P_0(D|\mathrm{Graph}_0)$. 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. To fix this problem, 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 $\lambda = \frac{1}{2}$). 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:

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

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

$$\mathrm{JSD}(P_1||P_0) = \lambda D_{KL}(P_1||M) + (1-\lambda) D_{KL}(P_0||M). \tag{4}$$

This quantity, JSD, 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), or in more technical terms it measures the total divergence to the average or the Information Radius (IRad) and it is this interpretability which makes this quantity so interesting and meaningful. Also note that this measure, unlike D_{KL}, is symmetric when λ is set to $\frac{1}{2}$ (i.e. $\mathrm{JSD}(P_1||P_0)$=$\mathrm{JSD}(P_0||P_1)$) and when the base 2 logarithm is used we have that $0 \leq \mathrm{JSD}(P||Q) \leq 1$, $\forall$ P, Q. Additionally, if the square root of this quantity is taken, a metric is obtained so that the triangle inequality holds, in addition to the measure being symmetric and non-negative ($\mathrm{JSD}(P_i||P_j) = 0$ iff $P_i = P_j$) which are important properties, if the measure is meant to be used to compare distances between different belief systems (i.e. if it is to be used to 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 the two models:

$$\mathrm{CD}(P_1||P_0) = \sqrt{\mathrm{JSD}(P_1||P_0)}. \tag{5}$$

3.3 Cognitive Diversity

An extension to JSD, (the n-point Jensen-Shannon Divergence, or JSD_n), can be defined as:

$$\mathrm{JSD}_n(P_1, P_2, \ldots, P_n) = H\left(\sum_{i=1}^{n} \omega_i P_i\right) - \sum_{i=1}^{n} \omega_i H(P_i), \tag{6}$$

where $H(\cdot)$ is defined as the Shannon Entropy and where the ω_is are weights that sum to 1. 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 come close to defining a collective's cognitive diversity as its current available theoretical tool kit and it seems that the JSD_n has exactly this quality; theoretical distributions are compared to a data sample and the more diverse these theories are (i.e. the higher is the numerical value of the JSD_n), the more theoretical material there is, among all of the models combined, to compare against the data. Note, however, that the maximum value of the JSD_n, which I may call the potential diversity, is an increasing function, of n, or more precisely, it is:

$$\text{Potential Diversity}(n) = \log_2(n), \tag{7}$$

which in the special case where $n = 2$ is equal to 1. In other words, it is a result of Equation 6, that $0 \leq \mathrm{JSD}_n \leq \log_2(\mathrm{n})$.

We can thus see that the potential diversity ($\log_2(n)$) is an increasing and concave function of the number of individuals, as intuition would suggest. Further, it is the case that for any tuple of probability distributions, for example the two-tuple (P, Q), the JSD_n over just that tuple has the same numerical value as JSD_{nk}, over any multiple of this tuple (k times the same entries as in the original tuple, where k is a positive integer). For example $\mathrm{JSD}_2(P, Q) = \mathrm{JSD}_4(P, P, Q, Q)$. Taken together, Equation 7 and this last point represent a problem for a measure of diversity, as having two, very different view-points in a collective of two people seems to mean, intuitively, that the collective of two is very diverse, while having two even radically different view-points in a collective of a hundred people seems to be decidedly less diverse. But luckily, there is an easy fix; devide the JSD_n, by potential diversity $\log_2(n)$ so that the final diversity measure is:

$$\text{Cognitive Diversity}_n(P_1, P_2, \ldots, P_n) = \sqrt{\frac{1}{\log_2(n)} \mathrm{JSD}_n(P_1, P_2, \ldots, P_n)}. \tag{8}$$

The cognitive diversity as defined in Equation 8, as it is normalized for group size, allows to compare the cognitive diversity of collectives with different numbers of individuals, as well as with the same fixed number of individuals, which was one of the central goals of this paper. 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, but rather to ask questions related to a group's diversity, relative to how diverse it could be.

4 Causal Explanations of the 2008 Foreclosure Crisis

Since the economic crisis[7] 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[8]. The material, alongside some explanations of how it is used to construct the cognitive maps, exhibited in Figure 2, can be found in the Appendix, although I shall give one example here: In the Financial Crisis Inquiry Commission Staff audiotape of the interview with Warren Buffett on the 26th of May, 2010, the Berkshire Hathaway CEO 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 ...

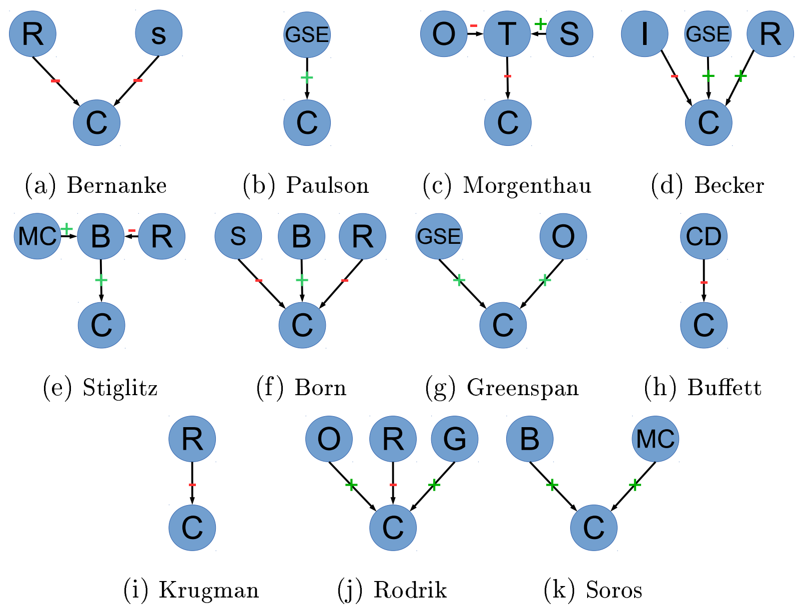
As Mr. Buffett only spoke of one cause, I will humerously name it CD: Cognitive Diversity, (virtually everybody succumbed) and as Buffet blamed the onset of the crisis on the lack of CD, I arrive at Mr. Buffet's CM (Figure 2h).

4.1 The Cast of Characters

- Ben Bernanke (economist and chairman of the US Federal Reserve Bank, Figure 2a),
- Henry Paulson, Jr. (past CEO of Goldman Sachs and Secretary of the US Treasury at the time of the crisis, Figure 2b),

[7]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.

[8]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 2: 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 2c),
- Joseph Stiglitz (economist, Figure 2e),
- Brooksley Born (Commissioner on the Financial Crisis Inquiry Commission and past chair of the Commodity Futures Trading Commission, Figure 2f),
- Alan Greenspan (economist and past chairman of the US Federal Reserve Bank, Figure 2g),
- Warren Buffett (CEO and largest shareholder of Berkshire Hathaway, known as the "Oracle of Omaha", Figure 2h),
- Paul Krugman (economist, Figure 2i),
- Dani Rodrik (economist, Figure 2j)

and

- George Soros (Chairman of Soros Fund Management and philanthropist, Figure 2k).

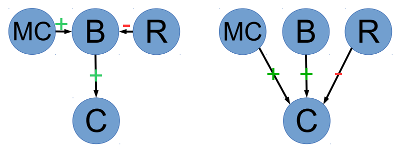
4.2 Cognitive Distances

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, $\sqrt{\mathrm{JSD}_2(i,j)}$, 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 2e) and Gary Becker (Figure 2d) are present in three out of the five largest diadic distances, while that of Robert Morgenthau (Figure 2c) 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 2k). 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 exag gerate 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 2a) and Brooksley Born (Figure 2f), 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 2j) and Brooksley Born (Figure 2f), 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 3: 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 3), 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
$n = 10$	-0.0112	-0.0055	0.0054	-0.0018	0.0058	-0.0091	-0.0107	-0.003	-0.006	-0.007	-0.0099
$n = 9$		-0.0078	0.0049	-0.0036	0.0049	-0.0094	-0.0135	-0.0047	-0.0067	-0.0081	-0.0125
$n = 8$		-0.009	0.004	-0.0037	0.0027	-0.0125		-0.0073	-0.0099	-0.0094	-0.0161
$n = 7$		-0.0133	0.0015	-0.0049	0.0031	-0.015		-0.0111	-0.0145	-0.0126	
$n = 6$		-0.0206	-0.0016	-0.0037	0.002			-0.0175	-0.0192	-0.0116	
$n = 5$			-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 9). Morgenthau, Becker, Stiglitz, Buffett, Krugman and Rodrik survived the iterated deletion of diversity minimizing elements, for a maximally diverse group of 5.

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)[9]. There are two ways in which a maximally diverse group of, say 5, could be constructed from a group of 10: 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 9) 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 10):

$$\min_i \left(\sqrt{\frac{JSD_n(\Omega_n)}{\log_2(n)}} - \sqrt{\frac{JSD_{n-1}(\Omega_n \setminus \text{Graph}_i)}{\log_2(n-1)}} \right), \text{ for } n = 10, \dots, 6, \quad (9)$$

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).

$$\max_i \left(\sqrt{\frac{JSD_{\tau+1}(S_\tau \oplus \text{Graph}_i)}{\log_2(\tau+1)}} - \sqrt{\frac{JSD_\tau(S_\tau)}{\log_2(\tau)}} \right), \text{ for } , \tau = 1, \dots, 4, \quad (10)$$

[9]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.

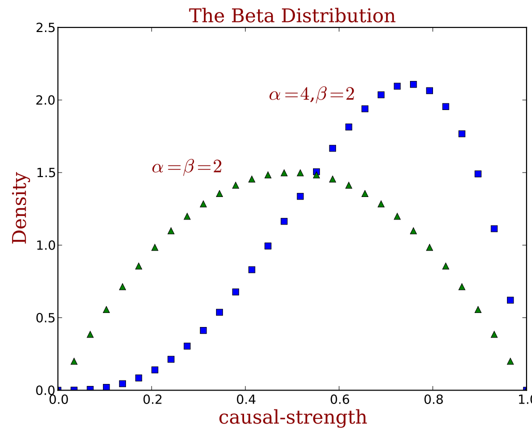


Figure 4: The beta distribution for two different parameterizations (α, β).

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 4). 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 \$$\alpha$\$s and \$$\beta$\$s, 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 has demonstrated how a theory of human causal learning, naturally gives rise to important information theoretic measures and how these 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 described and demonstrated a potential "hiring and firing" algorithm, if cognitive diversity were to be seen as a goal. I am hopeful that the tools that are here brought together and presented have potential to illuminate our understanding of collective thinking processes in the presence and absence of cognitive diversity. This paper was a first step.

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.

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.

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.