Topics in Cognitive Science 1 (2009) 598–620
Copyright © 2009 Cognitive Science Society, Inc. All rights reserved.
ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2009.01047.x

The Rules of Information Aggregation and Emergence of Collective Intelligent Behavior

Luís M. A. Bettencourt

Los Alamos Laboratory and Santa Fe Institute

Received 7 July 2008; received in revised form 14 May 2009; accepted 14 May 2009

Abstract

Information is a peculiar quantity. Unlike matter and energy, which are conserved by the laws of physics, the aggregation of knowledge from many sources can in fact produce more information (synergy) or less (redundancy) than the sum of its parts. This feature can endow groups with problem-solving strategies that are superior to those possible among noninteracting individuals and, in turn, may provide a selection drive toward collective cooperation and coordination. Here we explore the formal properties of information aggregation as a general principle for explaining features of social organization. We quantify information in terms of the general formalism of information theory, which also prescribes the rules of how different pieces of evidence inform the solution of a given problem. We then show how several canonical examples of collective cognition and coordination can be understood through principles of minimization of uncertainty (maximization of predictability) under information pooling over many individuals. We discuss in some detail how collective coordination in swarms, markets, natural language processing, and collaborative filtering may be guided by the optimal aggregation of information in social collectives. We also identify circumstances when these processes fail, leading, for example, to inefficient markets. The contrast to approaches to understand coordination and collaboration via decision and game theory is also briefly discussed.

Keywords: Collective behavior; Information theory; Cognition; Cooperation; Coordination; Natural language

1. Introduction

Decisions made by individuals and their aggregate consequences manifested at the level of organizations, from firms to nations, lie at the heart of all social and economic behavior

Correspondence should be sent to Luís M. A. Bettencourt, Theoretical Division, MS B284, Los Alamos National Laboratory, Los Alamos NM 87545; or Luís M.A. Bettencourt, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501. E-mail: lmbett@lanl.gov

(von Neumann & Morgenstern, 1944; Simon, 1957, 1991). The quest for quantitative theories of decision making motivated the birth of modern statistical science (Bernoulli, 1738), and it is a main driver of economic and social theory (Luce & Raiffa, 1957; Ross, 2007). With the increased richness of data in the social sciences made possible by the World Wide Web and the computational capacity to now model large-scale social behavior, the need for an integrated, actionable theory of collective behavior, connecting individual levels to social collectives, has never been greater.

From a practical point of view, the design and integration of simple models of decision making is increasingly important in synthetic cognition systems (Parsons, Gymtrasiewicz, & Wooldridge, 2002; Terano, Deguchi, & Takadama, 2003), such as in robotics, and in large-scale agent-based models, including those used to estimate the impact of large pandemics (Germann, Kadau, Longini, & Macken, 2006; Riley, 2007) and the management of critical infrastructure (Pederson, Dudenhoeffer, Hartley, & Permann, 2006). However, most approaches to these problems have relied on simple heuristics, specific to each problem at hand.

Decision theory and closely related game theory have been the standard microscopic starting points for models of human behavior and its computer modeling in terms of agent based simulations (Terano et al., 2003). Much criticism of their structure and assumptions has surfaced, questioning, for example, the plausibility of purely ‘‘rational behavior.’’ Bounded rationality (Simon, 1982) and a large number of decision heuristics that violate the expectations of decision and game theory (Camerer, Loewenstein, & Rabin, 2003; Ross, 2007) have in fact become commonplace, and new fields such as behavioral economics developed out of the shadow of their classical counterparts. Here, again, the lack of a principled approach, from which decision or game structures can be derived, has been lacking.

Several relatively recent technological developments are now increasing the need for a deeper understanding of the structure of decision making from deeper principles of information processing. In many important circumstances, such as online markets, collaborative filtering, and the dynamics of innovation of open source software communities (Tapscott & Williams, 2008), it is clear that real-time collective information signals are available to all participating decision makers. Changes in these signals feed back on human behavior and allow, in some circumstances, for the exploitation of information imbalances for profit or for enabling better solutions than those envisioned by classical rational decision makers (Tapscott & Williams, 2008). Thus, the collection and aggregation of information by individuals and groups can endow them with important advantages. The same is true in animal societies, from ants and bees (Hölldobler & Wilson, 2008) to social mammals (de Waal & Tyack, 2003).

In recent years the advent of large scale collaboration environments, enabled by the World Wide Web, is creating opportunities for new forms of information exchange and problem solving (Tapscott & Williams, 2008), and for their study by social scientists. However, the identification of the formal circumstances when a solution produced by an informal collective of individuals, each with partial information, can surpass in quality and speed those produced by experts or by dedicated organizations (Surowiecki, 2004) remains somewhat of a mystery. A better understanding of the individual decision processes in such environments and how they lead to collective solutions is a fascinating problem with multidisciplinary implications to the cognitive and other social sciences and economics.

Together these observations raise a few crucial questions, which will constitute the principal objectives of this paper. How is information formally aggregated across social scales from individuals to collectives and organizations? In what classes of problems can the pooling of individuals with incomplete information lead to a better solution than that of an expert individual or a purposeful organization? What are the structures of decisions and information sharing that can enable a collective to converge to a better solution than any of its parts? To address these questions we will establish a general quantitative framework where both general situations and specific problems can be classified in terms of their underlying information-processing strategies.

Before we start considering specific applications, we will have to understand how to deal with information quantitatively. Fortunately this is a standard issue in statistics and communications so that we will be able to borrow much of the formalism of information theory (Cover & Thomas, 1991). We will then pose the problem of ''collective intelligence'' in terms of the aggregation of the state of many statistical variables, each containing some information about a target variable *X*. These may denote, for example, the knowledge of several decision makers and how it is translated into an average population estimate. We show that there are general classes of problems and decision structures for which quantifiable collective intelligence is possible, while in others decreasing returns in the number of decision makers set in.

We then discuss how these insights may help build new forms of collaborative filtering and design better markets, by biasing the objective and the information available to individual users in ways that may foster collective intelligence. We shall also give some counterexamples where consensus is not desirable and where recommendation systems or efficient markets may prove impossible to implement. The remainder of this paper is as follows. Section 2 is the most formal, and it introduces several fundamental quantities in information theory and their mathematical properties. In section 3 we describe how information is aggregated across many variables, and we derive conditions for synergy and redundancy. We also provide a few examples of information aggregation. Section 4 discusses the application of these principles to several canonical examples of collective cognition. We show how several different problems of social organization may be understood from general principles of information processing in a unified way. Finally section 5, discusses several conceptual issues, as well as future challenges and applications.

2. The calculus of information

Information has had a long association with the cognitive sciences. The foundations of computer science were very much guided by cognitive considerations (Minsky & Papert, 1969; Turing, 1936; Wiener, 1948). Today, problems of coding, decoding, representation, and estimation, which are all based on formal aspects of information, play a role of growing importance in neuroscience (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997).

The cognitive sciences, while always inspired by models of computation, have arguably found less actual use for information theory (Luce, 2003), partly because of the difficulty of identifying and enumerating state spaces in, for example, psychology or the social sciences. A number of obstacles to the productive use of information theory are due to its best known role, in the theory of communications, where it is used to quantify the reliability of information transmission over noisy channels (Shannon & Weaver, 1949). Applied in this way to problems of cognition, such as the relation between sensory input and behavioral output, information theory can be helpful in characterizing ''codes'' that map one set of variables onto another (Miller, 1956), but this bypasses the most interesting processes in the mind. The language of information theory should become more fertile in the study of cognition when applied to other problems, related to the estimation of solutions and search, understood as the pooling of information necessary to find them. From this point of view, information theory allows us to assess the value of different individual pieces of information in solving a problem, as well as reveal the nature of the (generative) rules by which they make up the solution. It is in this later sense that we employ it here to reveal general principles of social cognition.

2.1. Quantifying knowledge uncertainty via Shannon's entropy

To measure how information is distributed in a population and to understand how it may be pooled optimally to produce coordinated collective behavior, we need to introduce a few standard concepts and quantities from information theory.

One of the most important facts to keep in mind is that information is a relative quantity, a sort of differential between two levels of uncertainty. Specifically, information refers to the reduction in uncertainty (or gain in predictability) of a given target variable, given knowledge of the state of another (Cover & Thomas, 1991). Mathematically, uncertainty can be cast in terms of the properties of the distribution of stochastic variables *X*. These variables may be fundamentally probabilistic (e.g., the result of a lottery), or, more frequently uncertainty about them arises from incomplete knowledge (e.g., the location of a food source). Whenever we can quantify the odds for the several states of *X,* we can quantify its uncertainty. It is most commonly expressed in terms of the Shannon entropy

$$S(X) = -\sum_{x} p(x)\,\ln_2 p(x), \tag{1}$$

where $p(x)$ is the probability associated with each state x of the variable $X,$ which we assume to be discrete for simplicity. We will see later that X may be a set of possible recommendations in collaborative filtering environments, or the price of an asset in a market, or the spatial location of a resource. Knowledge of the precise state of X means that $p(x) = 1$, for some specific state x, and consequently $S(X) = 0$, that is, there is no uncertainty remaining. Conversely, the entropy is maximal when there is complete uncertainty about the state of X, when $p(x)$ is the uniform distribution. Then, if there are N possible states, and $p(x) = 1/N$, the entropy is $S(X) = \ln\ _2N$. For example, if there are two equally likely choices, then the uncertainty is 1 bit; 0 or 1, or heads or tails. If there are four choices, it is 2 bits, and so on.

Mathematically, the entropy, and other information quantities to be discussed below, are *functionals* of the probability distributions of variables of interest, They take these functions and return a number that quantifies some property related to how well the solution of a given problem is known, or how much better we know it after a new observation of a related quantity.

2.2. Correlated variables and information aggregation

We will be interested in situations where the state of X can be gleaned (at least partially) from the measurement of other variables $\{Y\}_k$, generally correlated with X, and with each other. We adopt the notation for the set of k such variables $\{Y_1, Y_1,\ldots, Y_k\} = \{Y\}_k$. The variables $\{Y\}_k$ may be opinions in a collective, a history of product recommendations, or buy and sell price estimates for market assets. For example, if we knew the opinions of several traders about the price of an asset, how much could we tell about the real market price?

In other words, how does the pooling of the Y_i, inform us about the state of X? The information gain in the state of X, given observation of any Y_i, is also a familiar quantity in information theory (Cover & Thomas, 1991), known as the *mutual information* between the two variables:

$$I(X;Y_i) = S(X) - S(X|Y_i) \equiv -\frac{\Delta S(X)}{\Delta Y_i}, \tag{2}$$

where the first relation expresses the change in uncertainty of X, given knowledge about Y_i and the second is a definition, which reminds us that the information is a differential quantity, a change in uncertainty (entropy). This definition allows us to explicitly account for the measurement of variable Y as entropy reductions. It introduces a discrete calculus of information under variable conditioning, which generalizes to any number of variables, as we show below. Other information theory quantities are sometimes also useful in quantifying other aspects of information; see, for example, Nelson (2005).

The usefulness of the interpretation of information as a variation of the entropy of X under conditioning on Y_i will become more apparent below. Both the Shannon entropy and the mutual information between X and a set of variables $\{Y\}_k$ have important properties, namely

$$S(X|\{Y\}_k) \leq S(X|\{Y\}_{k-1}) \leq \cdots \leq S(X) \tag{3}$$

and

$$I(X;\{Y\}_k) \geq I(X;\{Y\}_{k-1}) \geq \cdots \geq I(X;Y_1). \tag{4}$$

These relations express the fact that measurement of successive variables Y_i can only increase (or at least leave unchanged) the information about X. We shall see below that, while being a natural property of information, these conditions help define classes of well-posed problems, where the search for the most information about X can be optimized in terms of the combination of the $\{Y\}_k$.

2.3. The rules of information aggregation

We are now ready to start unraveling the pattern of aggregation of information. Specifically we want to quantify how knowledge of many variables $\{Y\}_{k,}$ can inform the state of the target X. As we started to write above—for a single measurement—the information gained from a new variable Y_i, is computed by successively conditioning the current state of X on it. In order to do this generally, we write the information gained from measuring a set $\{Y\}_k$, which is (Bettencourt, Gintautas, & Ham, 2008)

$$\begin{aligned} I(X;\{Y\}_k) &= S(X) - S(X|\{Y\}_k) \\ &= -\sum_{i=1}^{k} \frac{\Delta S(X)}{\Delta Y_i} - \sum_{i>j=1}^{k} \frac{\Delta^2 S(X)}{\Delta Y_i \Delta Y_j} - \cdots - \frac{\Delta^k S(X)}{\Delta Y_1 \cdots \Delta Y_k} \end{aligned} \tag{5}$$

where the second and higher variations are obtained from the first via a chain rule, analogous to taking successive ordinary derivatives, for example,

$$\begin{aligned} \frac{\Delta^2 S(X)}{\Delta Y_i \Delta Y_j} = \frac{\Delta}{\Delta Y_j}\left[\frac{\Delta S(X)}{\Delta Y_i}\right] &= S\big(X|\{Y_i, Y_j\}\big) - S(X|Y_i) - S\big(X|Y_j\big) + S(X) \\ &= I(X; Y_j|Y_i) - I(X; Y_j) \end{aligned} \tag{6}$$

and so on. Note that the effect of the variation is to condition the information on the new variable and subtract the unvaried element from it.

Expansion (5 and 6) allows us to quantify how information is aggregated as more of the Y_i, are available, and, depending on their instantiation, can be mapped to many classes of problems. It is a useful decomposition in at least two ways. First, each term in the expansion refers to the irreducible information contribution of measuring a unique set of the $\{Y\}_{i<k}$. This contribution will vanish if any of the variables in the set are statistically independent, and thus do not form a genuine irreducible multiplet. Second, the sign of each term indicates how the set of the $\{Y\}_i$ variables are related to each other (and to X). They may reveal more information (if their sign is negative) or less information (if positive) than when organized at lower orders. It is in this very specific sense that we can say that knowledge of a set of variables may yield more insight than the sum of the information in its parts. This is the quantitative basis on which ‘‘collective intelligence’’ can be measured, as we explain in detail in the next section.

3. Collective problem solving: When is the sum greater than the parts?

In the previous section we developed the formalism necessary to quantify under what general conditions aggregating opinions, measurements, and other types of partial or imprecise information leads to greater information gain than that from the sum of each contribution relative to the target. Here we translate these conditions into general properties of probability distributions. This will allow us to make contact with a variety of specific problems in collective cognition in section 4.

3.1. Formal relationships between pieces of information and their aggregation

First, let us start with three elements: *X, Y₁, Y₂*. The information about *X* from the other variables is, from Eqs. 5 and 6,

$$\begin{aligned} I(X;\{Y_1,Y_2\}) &= S(X) - S(X|\{Y_1,Y_2\}) \\ &= -\frac{\Delta S(X)}{\Delta Y_1} - \frac{\Delta S(X)}{\Delta Y_2} - \frac{\Delta^2 S(X)}{\Delta Y_1 \Delta Y_2} \qquad , \\ &= I(X;Y_1) + I(X;Y_2) - R(X;Y_1;Y_2) \end{aligned} \tag{7}$$

where we used

$$\begin{aligned} &\frac{\Delta S(X)}{\Delta Y_1} = -I(X;Y_1); \qquad \frac{\Delta S(X)}{\Delta Y_2} = -I(X;Y_2)\,, \\ &R(X;Y_1;Y_2) \equiv \frac{\Delta^2 S(X)}{\Delta Y_1 \Delta Y_2} = I(X;Y_1) + I(X;Y_2) - I(X;\{Y_1,Y_2\}) \\ &\qquad = I(Y_1;Y_2) - I(Y_1;Y_2|X). \end{aligned} \tag{8}$$

Here we introduced the notation *R* for the second variation, for simplicity, which is also known as the coefficient of redundancy (Bettencourt, Stephens, Ham, & Gross, 2007; Schneidman, Still, Berry, & Bialek, 2003). We see that the information gain about *X* from knowledge of the pair Y_1, Y_2 is given by the sum of their independent mutual information with *X* (the first-order variations), and a correction, which accounts for the effects of the correlations between the two *Y* variables. This correction is the difference between the information shared between Y_1, Y_2 given knowledge of *X*, and the information that the two Y_i's contain about each other, regardless of *X*.

This expression shows when there is a benefit to pool information between the Y_i's, and when there is not. These two situations correspond to when the second variation is negative (synergy) or when it is positive (redundancy), respectively. Specifically if the Y_i's are mutually independent, which in terms of probabilities means

$$P(Y_1,Y_2) = P(Y_1)P(Y_2), \tag{9}$$

then the second term in the last line of Eq. 8 vanishes. Note that

$$I(Y_1;Y_2) = \sum_{y_1,y_2} P(y_1;y_2)\log_2\left[\frac{P(y_1;y_2)}{P(y_1)P(y_2)}\right]),$$

and it always pays to share information, given that the Y_i's are correlated to the objective *X*.

Conversely if the Y_i's are conditionally independent, given X, which means that

$$P(Y_1, Y_2|X) = P(Y_1|X)P(Y_2|X), \tag{10}$$

but not mutually independent, then the Y_i's are at least partially redundant and pooling information results in less than the sum of the individual lower-order terms. That is, to say, in this situation there is at least partial redundancy between the information contained in the Y_i's.

Relations (9 and 10) generalize to k variables, with mutual independence written as

$$P(\{Y\}_k) = P(Y_1)P(Y_2)\cdots P(Y_k), \tag{11}$$

and conditional independence as

$$P(\{Y\}_k|X) = P(Y_1|X)P(Y_2|X)\cdots P(Y_k|X). \tag{12}$$

These conditions are sufficient to classify problems where collective coordination is advantageous. The optimal requirement is simply that each contribution is statistically independent from others and that they are *not* conditionally independent given the state of the target X. Below we show that in most problems of collective cognition there is a natural aggregator of information $X = f(\{Y\}_k)$, where f is a general function, that makes synergy possible.

Note that there are two separate ingredients contributing to the possibility of an optimal synergetic strategy: (a) the fact that the information aggregator X does not create conditional independence of the several contributions, which makes synergy possible, and (b) that given the possibility of synergy, each component remains as independent as possible from the others. The design of a good aggregator is the result of individual cognitive processes, or of the way social signaling, for example, a market or recommendation system, is prescribed. On the other hand, the independence of information gathering is typically the result of behavioral or decision choices on the part of the social participants in the system. There are actionable insights to be drawn at both these levels, which we discuss in section 5.

3.2. Four examples

Let us first show how this works with a few familiar examples. In the next section we analyze more complex examples such as markets, recommendation systems, and parallel searches, implemented by social insects and in several agent-based models.

3.2.1. Guessing the weigh of an ox

Consider first a simple set of games where several players need to guess the answer to a given question: For example, the opening anecdote in James Surowiecki's book *The wisdom of the crowds* (Surowiecki, 2004), where Francis Galton studies the outcome of guessing the weight of an ox at an English country fair (Galton, 1907). Each individual opinion Y_i may be thought of as an independent stochastic variable. The opinions of some people may have greater uncertainty than others', reflecting, for example, their knowledge and personality. What is the best way to use people's knowledge to estimate the weight of the ox?

We imagine first a situation where an expert Y_1 makes his guess. Because he trusts his own judgment above all, he chooses not to pay attention to the opinions of others. As a result the uncertainty remaining in X is $S(X|Y_1)$. This may indeed be a good guess but it does not use all information available: It is not, in this sense, optimal.

Next consider instead a somewhat less experienced individual Y_3, who will listen to the opinion of one of his friends Y_2. His opinion will completely reflect his friend's opinion so that, conditional on this knowledge, no new information is gained, that is, $I(X;\, Y_3|Y_2) = 0$. As a result it follows that the information gain about the ox's weight X from both their opinions is

$$\begin{aligned} I(X;\{Y_2,Y_3\}) &= I(X;Y_2) + I(X;Y_3) + [I(X;Y_2|Y_3) - I(X;Y_2)] \\ &= I(X;Y_2). \end{aligned} \tag{13}$$

We see that the information provided by Y_3 is entirely redundant and does not contribute to solving the problem, as we might have expected. We could of course have conceived a situation where Y_3 listens to his friend only partially, still contributing with some new information to the determination of X, but less than if his information were completely independent of that of Y_2. In such circumstances the solution would still benefit from Y_3's contribution, but less so than if he would have made up his mind on his own.

Finally consider Galton's approach (Galton, 1907; Surowiecki, 2004) of considering all opinions in the crowd to be equally valid by averaging over all k participants, so that the estimate for the weight is

$$X = \frac{1}{k}\sum_{i=1}^{k} Y_i.$$

Assume, for simplicity, and in contrast to the previous case, that the many Y_i do make up their minds independently so that for any Y_i and Y_j $I(Y_\mathrm{i};\, Y_\mathrm{j}) = 0$. Then, because the average that defines Galton's estimate for X, forces the variables to be conditionally dependent (it expresses X as a function of the Y_i), expressions (5 and 6) tell us that the resulting estimate must be *better* than the sum of the information from each individual, that is,

$$I(X;\{Y\}_k) > \sum_{i=1}^{k} I(X;Y_i). \tag{14}$$

This is a remarkable result: ''Collective intelligence'' can emerge from pooling independent information so long as the pieces are conditionally dependent, as designed by a mechanism of aggregation. The average works, of course, and is a natural function to produce an aggregate estimate in many circumstances where all variables refer to the same quantity, but in fact any joint function $X = f(\{Y_i\})$ would equally do the trick. Note that if individual opinions are not mutually independent, then it is still possible to gain knowledge about the end result by pooling in more people, but a larger number will be necessary to make up the same information.

3.2.2. Uncertainty in coupled Gaussian variables

The next example is very familiar in estimation, but it is not always necessarily thought of in terms on information. It also allows us to show that there is more to collective aggregation of information than the central limit theorem, which tells us the statistics of sums of stochastic variables given their individual statistics.

First, consider as before X as the mean of the Y_i

$$X = \frac{1}{k}\sum_{i=1}^{k} Y_i.$$

For simplicity and transparency let us also take each of the Y_i to be independent and identically distributed (iid) normal distributions so that

$$Y_i \sim N(\mu, \sigma^2); \qquad X \sim N(\mu, \tfrac{\sigma^2}{k}).$$

Clearly, although independent the Y_i are not *conditionally independent* as their sum (divided by k) must add up to X. Then it follows that they must be synergetic, as in Eq. 14. The uncertainty (entropy) in each of the Y_i is, using a well-known result for Gaussians distributions (Cover & Thomas, 1991),

$$S(Y_i) = \ln[\sqrt{2\pi e}\sigma].$$

By the same token, the initial uncertainty of X *is*

$$S(X) = \ln[\sqrt{2\pi e}\frac{\sigma}{\sqrt{k}}],$$

which is also the total information contained in the set $\{Y\}_k$ about X. We see that the central limit theorem implies that the uncertainty of the *average* becomes smaller and smaller as the number of Y_i variables increases. This statement relates the properties of the *statistics* of X to those of the $\{Y\}_k$. But what about the *states* of the several variables?

How is the state of X determined as we know more and more of the Y_i in $\{Y\}_k$? Clearly when we know all k Y_i, X is know exactly, and $S(X|\{Y\}_k) = 0$. When a number $0 < n<k$ of Y_i are known it is a straightforward calculation with Gaussians to show that

$$S(X|\{Y\}_{n<k}) = \ln\left[\sqrt{2\pi e\frac{k-n}{k^2}}\sigma\right]$$

This result is shown in Fig. 1, for $k = 2$ and 3: The important feature is that the curves accounting for the uncertainly left in X are convex (the second derivative relative to n is negative). In this sense, and despite each variable containing by itself the same amount of information about the average, knowledge of later variables reveals more information about its state than earlier ones. This is synergy.

3.2.3. Reverse-engineering logical circuits

Other examples of interesting information aggregators that cause synergy are logical circuits. Consider inputs $\{Y\}_{\mathrm{k}}$ of, say, an AND or a XOR, that are statistically independent variables, and the output X is the circuit's operation (Bettencourt et al., 2007; Schneidman et al., 2003). Just as for the average we have now that $X = f(\{Y\}_{\mathrm{k}})$.

Fig. 2 shows the results for a simple AND circuit with two independent random inputs. Circuits that are a function of k general inputs show synergy among larger sets of variables.

3.2.4. Markov chains and the flow of information

Finally, consider another example closer to models of sequential decisions: a Markov chain of events. A Markov chain of order k is a set of stochastic variables ordered in some specific way—for example, in a temporal sequence—where the state of a new variable to be measured depends only on a number of previous ones, and none before a certain order k. The simplest and most common example is $k = 1$, where the state of a variable X, depends only on the previous measurement Y_1 but not on Y_2, Y_3, etc. Formally this means that, for a Markov chain of order k

$$P[X; Y_{k+1}|\{Y\}_k] = P[X|\{Y\}_k]P[Y_{k+1}|\{Y\}_k]$$

As a result we see that X and $Y_{\mathrm{k+1}}$ etc. are conditionally independent, given $\{Y\}_{\mathrm{k}}$. They are, however, not independent as they are generated by the same Markov process. As a result

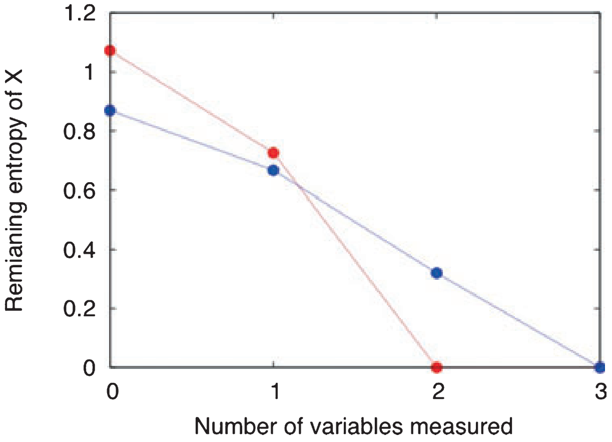
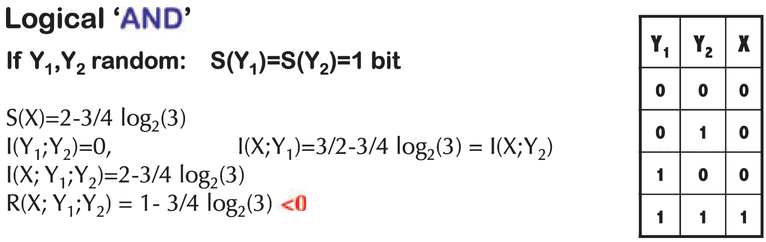


Fig. 1. Synergetic relations between conditionally dependent variables make information gain of successive measurements larger than expected from estimates based on the sum of mutual information $I(X;Y_{\mathrm{i}})$, Red dots show measurements for an average of $k = 2$ variables; blue for $k = 3$ variables ($\sigma = 1$).



Logical 'AND'

If Y_1,Y_2 random: $S(Y_1)=S(Y_2)=1$ bit

$S(X)=2-3/4 \log_2(3)$

$I(Y_1;Y_2)=0$, $I(X;Y_1)=3/2-3/4 \log_2(3) = I(X;Y_2)$

$I(X;Y_1;Y_2)=2-3/4 \log_2(3)$

$R(X;Y_1;Y_2) = 1- 3/4 \log_2(3) <0$

Y_1	Y_2	X
0	0	0
0	1	0
1	0	0
1	1	1

Fig. 2. Information quantities for an AND circuit with two random independent inputs X, Y. Note that $R < 1$, so that synergy of the inputs is possible in general.

their knowledge is entirely redundant (Bettencourt et al., 2007; Schneidman et al., 2003). In practice this means that, for $k = 1$, measuring Y_1 suffices to determine the state of X, and that other measurements yield much less information gain (in fact none at all) than expected based on their correlation with X alone. The result generalizes as expected to order k, where only the first k Y_i need to be known.

4. Searches, decisions, and the informational dynamics of aggregates

Now that we have laid down the formal relations to reason in terms of information aggregation let us revisit some familiar examples of collective information pooling, and see what category—synergy or redundancy—they may belong to.

4.1. Collective spatial searches and foraging

First let us consider the problem of collective spatial searches. Often, in the literature of complex systems, these problems are posed as the identification of a set of behavioral rules inspired by the organization of social insect colonies. Similar situations arise in the study of flocks of birds or schools of fish (Couzin, 2009). A variety of algorithms based on a set of more or less accepted heuristics have been developed to model the process of collective search such as ''ant colony optimization'' (Dorigo & Stützle, 2004) or the ''bees algorithm'' (Bonabeau, Dorigo, & Theraulaz, 1999). Although useful in some instances, the circumstances where these interesting approaches work or fail has remained obscure.

In general terms the search goes as follows. Several insects depart from a starting point—a nest—and explore space in order to find food, or a new nest, the reward. Collective coordinated searches often lead to good algorithms. For example, ants that find food quickest will return to the nest first, laying down a trail of pheromones that others can follow. This process will reinforce exploitation of discovered paths if any ant at the nest has a tendency to follow the strongest pheromone track, which is usually the first path to food to be found. Bees work in similar ways. Individual bees explore space around the nest and, upon returning, perform a ''dance'' that is in its orientation and duration indicative of the location and quality of the resource. Often several bees will return with information about several targets, which is conveyed to others through dances. After a number of trips, informed by these dances, the colony usually converges on a consensus target (Seeley & Visscher, 2004). It is sobering that individuals so simple—in the words of Bert Hölldobler, and Edward O. Wilson ''One ant alone is a disappointment'' (Hölldobler & Wilson, 2008)—can solve problems of such extraordinary complexity. The magic lies of course in the role of acquired information in selectively coordinating social behavior toward an optimal, or at least good, decision. But how is this achieved formally?

There are actually two aspects to this strategy, one concerned with exploration and another with exploitation of acquired information. First, before resources are found, several searchers will follow individual paths in search of promising locations. In traditional algorithms this search can be performed at random, although as we shall see below, given clues about the location of resources, aggregation of information can—in general circumstances—lead to cooperative coordinated behavior among searchers. Secondly, once found, paths to food can be exploited by other agents, under the assumption that no further information is necessary to forage additional resources. This confers greater fitness to the agents, despite no new information being uncovered. This example thus clarifies the balance between exploration (finding a path to the resource) and exploitation (using the best paths) in conferring reward, as usually expressed in the context of reinforcement learning (Sulton & Barto, 1998).

The first kind of behavior—exploration—must be preferred when there is large uncertainty about where the reward may be found X, whereas the second—exploitation—should follow when little doubt exists about the best location, and consensus in the colony has been achieved. The balance between exploration and exploitation can be captured in terms of an information theoretic objective function F

$$F = S(X) + \lambda C(X; \{Y\}_k). \tag{15}$$

Here $S(X)$ is, as before, the uncertainty in the position in space X of the resource, whereas $C(X; \{Y\}_{\mathrm{k}})$ is the cost in performing ''measurements'' at positions $\{Y\}_{\mathrm{k}}$, which can be negative (but bounded from below), expressing reward. The parameter λ establishes the balance between the two types of behavior, with exploration dominating for small λ. At each point in time the best actions Y_i can be judged by minimizing Eq. 15. The process of pure exploitation is straightforward. It can be expressed in terms of classical decision theory as the action that maximizes utility (reward) in the absence of risk (Luce & Raiffa, 1957) (when there is no uncertainty, i.e., $S(X) = 0$). We shall concentrate below in the minimization of the first term, that is, the collective actions by a colony of agents that lead to the minimization of uncertainty about the location of the resource.

We start by formalizing the problem in terms of a probability density $P(X)$, quantifying the ab initio knowledge of where rewards may be found. If no information is known, then this distribution is uniform in space. If no clues about the location of the reward are available at other locations, then any searcher should explore space randomly and the problem is trivial. More interestingly we will assume that the location of the reward is a source of a stochastic process that emits clues that are to be found at position Y, with some probability

$P(Y|X)$. Typically these ''clues'' are less likely to be found if the searcher is farther away from the source location. The formulation of this problem—for a single searcher—was elaborated recently by Vergassola, Villermaux, and Shraiman (2007). Clues detected at different positions Y_1, Y_2, …, Y_k are aggregated to successively update $P(X)$, and this new information is then used to perform an optimal search for the source position. This indeed gives the most general way to aggregate information via Bayes' theorem, specifically

$$P(X|Y_k) = \Lambda P(Y_k|X)P(X) \tag{16}$$

where Λ is a normalization factor. Successive measurements Y_i can be integrated in this way through repeated applications of Bayes' theorem, by using the posterior distribution on the left-hand side of Eq. 16, as the next prior $P(X)$ on the right-hand side. Note that as more measurements are performed the properties of information (3 and 4) guarantee that the uncertainty in the position of X must decrease or stay constant. It will only stay constant if the measurements performed by the searcher over time are totally redundant with each other. A strategy for optimal search can then be devised by taking steps, that is, measurements at a point Y_i, so that

$$Y_i : \quad \frac{\Delta S(X)}{\Delta Y_i} = \min\left[\frac{\Delta S(X)}{\Delta Y_j}\right], \quad \forall_j$$

A more interesting solution is possible when several measurements can be performed simultaneously by a number of agents. Then we face the situation where expansion (5 and 6) applies, and where coordination between searchers is possible in general.

As we know by now, synergetic searches—which coordinate the moves of searchers to maximize total information gain—are possible if measurements are not conditionally independent, that is, if the several measurements interfere—negatively or positively—with each other. This is usually a consequence of the measurement at a given point affecting the probabilities of detection at another. In social insects this is probably implemented via the signaling between individuals, which is dependent on the quality and location of the resource, and leads to the recruitment of new individuals to explore each source. In this sense the swarm can act much more intelligently as a whole by pooling information more quickly than the sum of its parts, a realization of course with important consequences for evolutionary theories of socialization.

Interestingly, these conditions are also implicit in the search problems faced by groups of predators and by a swarm of prey. The predator is faced with the decision problem of singling out a prey among many. Conversely, prey seek to maintain uncertainty of motion high for the faster predator. Motion from a group of predators that reduces this uncertainty by localizing the prey is analogous to the abstract search process described above. Reaction from the swarm of prey to stay and move together maintains the uncertainty from the point of view of the predator and may defeat its moves guided by the intent to reduce uncertainty and zero in on an individual. In this sense, the collective motion of predators and prey may be understood as the result of general principles of information management, which, upon analysis of each species in the presence of the other, may allow us to derive classes of rules of behavior for swarms, so far prescribed on purely empirical grounds (Bonabeau et al., 1999; Couzin, 2009; Dorigo & Stützle, 2004). This would be an important achievement of the framework proposed here.

4.2. Information and the structure of natural language

With sustained exponential increases in text collection and the power to process vast amounts of data, several areas of feature classification and prediction are increasingly less limited empirically. Two main areas of cognitive sciences are benefiting primarily from these large datasets: vision and natural language.

Natural language has always been a main focus of research in cognitive sciences (Chomsky, 1957; Pinker, 1999). Many of the semantic considerations of sentence interpretation and sentence construction remain fascinating but difficult problems. While these issues certainly hinge partially on information structures, we will not attempt to address them here. The processing of written natural language is an interesting developing field as text is easy to harvest and the structure of words and sentences gives a clear target for prediction. In fact, increasingly, almost every book, scientific paper, as well as blogs, and all kinds of formal and informal written text can be harvested online. The result is a new set of very large language corpora. For example, the Google *n*-gram set released in 2006 (Brants & Franz, 2006) contains over a trillion tokens (mostly words), from publicly available Web pages across the Internet. Other large—and cleaner—datasets, such as those derived from Wikipedia (2.3 billion tokens, 7 million unique) or project Gutenberg (about 15,000 books in English; 1.1 billion tokens, 1.9 million unique) contain billions of word occurrences. What breakthroughs in linguistics and natural language processing can be achieved using these resources?

Some of the technological improvements made possible are already familiar from text prediction and spell checking online, now automatically performed by Google and other companies. The existence of very large corpora of natural language should also facilitate efforts to train synthetic cognitive systems for natural language processing, possibly testing in this way hypotheses about the viability of specific mechanisms of linguistic cognition in humans. In this sense corpora of written language far outstrip in size the resources available to train computer vision systems, for example.

Here, to illustrate some of straightforward uses of information theory in natural language processing we focus on still interesting but simpler questions of predicting word sequences. For example, given the first few words of a sentence, how well can one predict what words will follow? The first attempts to tackle this sort of question were due to Shannon himself (1950), who studied character sequences in written English.

Effective word prediction suffers from the familiar combinatorial explosion resulting from the vast numbers of semantically and syntactically valid sentences that can be created through the combination of tens or hundreds of thousands of words available to most speakers. This effect is demonstrated in Fig. 3, where we show the percentage of unique *n*-grams (sets of consecutive *n* words) in written English as a fraction of all occurrences in the

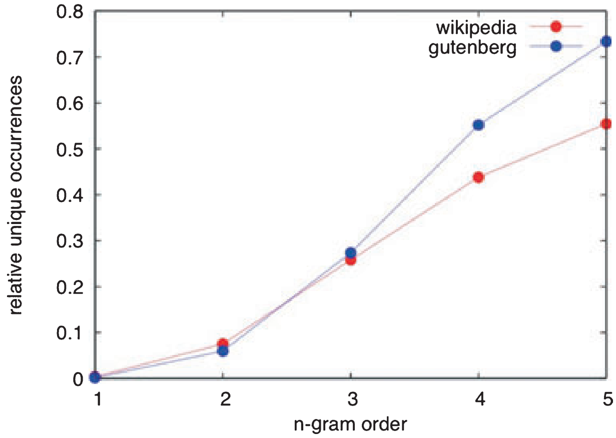


Fig. 3. The fraction of unique n-grams of varying size for Wikipedia and Project Gutenberg. By about $n = 4$ half of all n-grams are unique, exposing the generative nature of natural languages to produce very rare but perfectly intelligible ''new'' sentence structures.

Gutenberg and Wikipedia corpora. By about $n = 4$, half of all n-grams is unique, that is, it occurs only once in the corpus.

This means that as n gets larger than about four or five, most sequences of words cannot be predicted on the basis of their frequency, even in extremely large corpora. Nevertheless n-gram models may be useful to predict short word sequences. These models frequently make the assumption (Brown, Mercer, Della Pietra, & Lai, 1992) that word occurrence can be predicted on the basis of a Markov chain model of order n; see section 3.2.4 above. These models are written in terms of the conditional probability of a word given a preceding sequence of other terms in a sentence Y_{i}

$$P(X|Y_1, Y_2, \cdots, Y_n).$$

Such probabilities are estimated from frequencies of occurrence in text, or via Bayesian data assimilation schemes, just as we described above, *mutatis mutandis*. Many refinements to this picture and strategies of estimation of the n-gram sequences exist in the literature and will not concern us here (Manning & Schütze, 1999).

What interests us is whether a sequential strategy, based on predicting the next word in a string of tokens is warranted. It is well known from the study of other cognitive processes, especially vision, that elements of the same scene are processed in parallel. Is there a cognitive advantage to processing language also in this way, assembling different words in a sentence together simultaneously? The answer is yes, of course, if words (or word combinations) are synergetic. If so, together they would confer greater meaning to the sentence than when taken separately (see e.g., McElree, Frisson, & Pickering, 2006), or in a sequential manner that does not re-evaluate word occurrence patterns in a sentence.

Is language synergetic? As we know by now, this depends on the nature of the correlations between the elements of the *n*-gram Y_i. The answer is that it depends on the specific word or word sets under consideration. Not all words and word combinations are equally predictive of the next word *X*. Many of the Y_i are certainly not statistically independent, regardless of *X,* that is, they occur in predictable combinations regardless of the words that follow. Examples are ''as well as'' or ''in spite of.'' Such word combinations are frequent but have little predictive power for what follows. Secondly, it is the combinations of words that are maximally conditionally dependent given *X* that are the best predictors of the next words. These may be words in sequence, or they may in some cases exist in the sentence after the occurrence of the target word, as in a ''if … then'' clause. Fig. 4 shows words *X* (the, he, she, dog, cat and pug, in order of their frequency), and their corresponding coefficient of redundancy, *R*. This expresses how predictive each of these words is of combinations of two other words that follow. We see that frequent and general-purpose words like ''the'' have very little predictive power—are redundant—whereas rarer words such as ''dog,'' ''cat,'' and ''pug'' are indeed predictive of group of words that follow. Thus, some words do not constitute a good basis for prediction of sentence structure and are in fact empirically included in stop-lists to be excluded from statistical analysis. Others, however, are predictive of the context that surrounds them, expressed in terms of groupings of words. These words constitute candidates for word prediction. Words like ''the'' or ''a'' though not predictive are of course quite *predictable*, given sequences of other words nearby. This emphasizes how language is most likely to not be processed sequentially but by a combination of semantic and grammatical constraints anchored on a few key terms that determine structures around them.

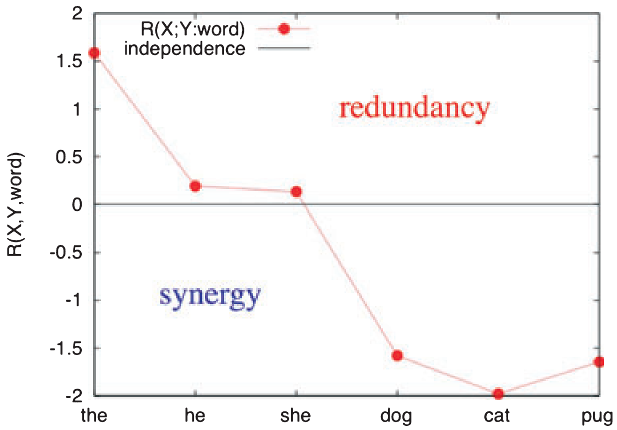


Fig. 4. Syntactic structures constrain predictable word occurrences but different words imply the presence of others to very different degrees. The presence of ''the''—the most frequent word in English—is not predictive of the words that follow or of their combination (Gutenberg Project corpus). Rarer words, such as ''dog,'' ''cat,'' and ''pug,'' are more predictive. These features result in redundancy or synergy, respectively expressed as positive or negative *R*. Rarer words tend to be more synergetic.

Analyzing a sentence in these terms is more analogous to thinking of it in terms of a logical circuit than of a Markov chain. This perspective may be a more appropriate formal model for language prediction, as we argued here based on general information theoretic considerations. Such circuits can indeed be reverse engineered (Bettencourt et al., 2008) given sufficient frequency data and it will be interesting to explore to what extent they may provide better models of language identification and prediction. The resulting structures would be ontologies, learned in an unsupervised manner, that also capture equivalent classes of groups of words under the same semantics. It would be interesting to compare such structures, retrieved automatically via the minimization of information theoretic objective functions, to sentence structures singled out by cognitive linguists (Dennis, 2004; Jones & Mewhort, 2007).

4.3. Information and the efficiency of markets

Markets are ubiquitous in human societies. They not only determine the price of most assets and commodities worldwide today, but they are also increasingly used for prediction of outcomes, such as elections, product introductions, and of course the results of sporting events (Surowiecki, 2004; Tapscott & Williams, 2008).

Markets provide an (usually financial) incentive for any investor with information that impacts prices to exploit it in order to make a profit. One of the cornerstones of financial theory is the efficient markets hypothesis (EMH) (Fama, 1970). This says essentially that markets adjust instantaneously to new information to correctly price assets. Much debate—both academic and ideological—has accompanied tests and assumptions of this hypothesis, including in important issues of policy. Curiously, the evidence is somewhat checkered. In many, perhaps most, circumstances markets can predict objective events with great precision, often better than expert opinions or polls (Surowiecki, 2004). On the other hand, evidence for famous investors systematically outperforming the market and the occurrence of bubbles and busts point to times and places where markets fail to reflect true value (Surowiecki, 2004). The observation of individual irrational biases in behavior such as framing of decisions, anchoring of prices to arbitrary positions, herding in decision making, a bias toward loss aversion, etc. (Ariely, 2009) are examples where rational decisions are foregone by individual investors. In this sense individuals exhibit at best bounded rationality (Ariely, 2009; Camerer et al., 2003; Ross, 2007; Simon, 1982) and often fall back on decision heuristics under uncertainty or cognitive overload.

Which of these behaviors at the individual level may lead to inefficient markets? Or can they somehow be exploited and corrected immediately by efficient markets? Here we rephrase what a market is, in terms of the information quantities introduced above, and explore some of their formal consequences.

Markets produce prices for listed commodities and assets. The price of anything at any given time corresponds to where offers to sell and offers to buy meet each other. In this sense price X is the aggregator of information, from many traders $\{Y\}_{\mathrm{k}}$ all with different motivations and opinions. The fact that X, may not result from Bayesian data assimilation or from the averaging of opinions is, as we discussed above, immaterial. It binds together the offers of buyers and sellers to produce conditional dependence between their values. From this perspective the price of an asset is *potentially* a synergetic indicator of the information manifested by all active traders in terms of their offers to buy and sell. Thus, the general structure of markets certainly has the potential for aggregating information efficiently. The Achilles heel of markets resides, however, on the sources of correlation between traders.

Specifically, what is missing from these considerations to provide an optimally efficient market is the independence of bids Y_i among traders, as well as the same trader over time. Cognitive biases and the application of the well-known decision heuristics mentioned above certainly destroy the statistical independence between traders' actions and may render markets inefficient (see also Salganik and Watts, this volume). Herding, for example, correlates the actions of many traders. Momentum investing indicates that there is a dependence between decisions at different times, often by the same investors. Informed investors, more certain than most of the true value of an asset, may have no incentive to reveal their information instantaneously, especially if increasing temporary imbalances may benefit their strategy and bottom line. Similarly as news is often available via broadcast, many investors may be coordinated by external sources, resulting in artificial moves that are not statistically independent, and may move price temporarily beyond the impact warranted by the information content. Thus, any behavioral heuristic or external signal that destroys the independent decision making and actions of participants in a market can potentially destroy the efficiency of the market as a whole. Because such endogenous and exogenous coordinating mechanisms are ubiquitous, it is important to recognize that markets may be (temporarily) wrong in pricing assets or generating predictions, an observation that carries important economic and political consequences. The information theoretic considerations developed here provide the basis for specific statistical tests that can reveal the degree of efficiency of a given market, especially if individual information about traders' behavior is available.

4.4. Collaborative filtering and recommendation systems

Contrary to the examples above, where collectives—swarms, markets—can aggregate individual knowledge to produce information about a problem that is superior than the sum of the parts, collaborative filtering is based on the exploitation of redundancy of behavior and taste.

Collaborative filtering (Goldberg, Nichols, Oki, & Terry, 1992) refers to a set of techniques and practices to aggregate information among a set of users to produce new recommendations, usually about a product or service that they may consider adopting or purchasing. The essence of the approach is to characterize a user in terms of a product set, or other characteristics, and through similarity with other users (or product sets) generate new recommendations, likely to be accepted.

Consider then $P(X)$ as the probability density over a discrete space of products that can be adopted by the user, for example, books or video rentals. In the absence of collaborative filtering, a recommendation may be made using products associated with the most probable states of this distribution. Typically the uncertainty remaining, which can be expressed in terms of $S(X)$, is large and it is desirable to include information from the preferences or product history of other users $\{Y\}_{\mathrm{k}}$, in order to minimize $S(X|\{Y\}_{\mathrm{k}})$. But the set of other users may be enormous, so what then is the procedure to choose the Y_{i} that best predicts the tastes of X?

Formally this problem is identical to the spatial search above. Expansion (5 and 6) does indeed tell us that the best single user Y_{i} is the one that has the most similar taste profile to X: Pick Y_{i} such that

$$Y_i : \qquad \max I(X;Y_j), \quad \underset{j\in\{1,..,k\}}{\forall},$$

that is, the user preference distribution with the highest mutual information with X. This is usually what is implemented—not necessarily in terms of mutual information—in typical recommendation systems.

Incorporating the preferences of several users takes us to different, although by now familiar, territory. If some of the history of adoptions by X and the set $\{Y\}_{\mathrm{k}}$ are available (e.g., as a time series of purchases or adoptions for all users), then we can estimate both the joint distribution with groups of other users, and the marginals for each user. With this information in hand, we can now weed out redundant users in the set $\{Y\}_{\mathrm{k}}$ and pick the smallest subset of users (and their preferences) that most certainly determine the preferences of X. As a result we will typically refrain to use a set of $\{Y\}_{\mathrm{k}}$ that contains users very similar to each other because they are not mutually independent. A similar procedure can be used to reverse engineer the ‘‘circuits’’ that predict the probabilistic states of any node in a general network, given the observation of the states of others with whom the former is correlated (Bettencourt et al., 2008).

5. Discussion and outlook

Social ensembles often display remarkable aggregate abilities to solve complex problems, more quickly and efficiently than experts or dedicated organizations. In other circumstances, collectives may amplify irrational behavior. Are there general principles that allow us to conceive in unified ways, across sets of applications, the circumstances and mechanisms when swarms, markets, or recommendation systems are capable of aggregating information in ways superior to the sum of knowledge in each of their components? Or when they are likely to fail?

Here we have suggested that the best starting point to answer this question is to understand how information is aggregated among individuals with diverse knowledge and opinions to generate a sum total of the social collective. Information in this respect is a peculiar quantity. Unlike matter or energy, it obeys a specific but often unfamiliar form of differential calculus as different observations are aggregated. As we have shown, information does not aggregate linearly, and it tends instead to generate more or less knowledge on the whole about a given problem X than the sum of the information in each part of the system.

All forms of aggregation—averages, market prices or Bayesian inference—rely on the pooling of information from different sources so that the target X and the observations $\{Y\}_{\mathrm{k}}$

are constrained together and become conditionally dependent. This is a necessary condition for synergetic aggregation of information, but it is not by itself sufficient. In addition, it is essential that the set $\{Y\}_k$ is ''sufficiently'' independent. The clearest case emerges when the set $\{Y\}_k$ comprises variables that are all statistically mutually independent, in which case any aggregation mechanism suffices to produce synergy. In real-life situations it is usually enough that sufficient independence exists, and predictive subsets of the $\{Y\}_k$ can be chosen that weed out redundant variables.

The fundamental problems in economics and the social sciences of the emergence of cooperation, consensus formation, and ''collective intelligence'' have been most often approached in terms of iterated non-zero sum games (Axelrod, 1984; Axelrod & Hamilton, 1981; Nowak, 2006), and other decision structures, which build in (unstable) incentives to coordination. These formalizations have generated precious insights into the cognitive mechanisms of individuals and social collectives. However, in real-life situations, people often do find novel ways of cooperating in collective problem solving even in the absence of explicit interaction rules or tangible rewards. The process of discovery of cooperating states, and the general dynamics that may enable them, has an important informational component that we argued here can be understood in general terms whether it concerns coordinated searches (Bourgault, Furukawa, & Durrant, 2003; Seung, Opper, & Sompolinsky, 1992), markets, or collaborative filters.

We have shown that *any* mechanism for the aggregation of information can lead—given enough independence of the parts—to synergetic interactions. Given that privileged knowledge equates to social (and often financial) advantage: Who gets to aggregate information? Clearly information aggregation is a ubiquitous process, central to learning. It can take place at many different levels from the individual, who may extract personal advantage from it, to social collectives, where aggregated information such as market prices become a public good, open to individual challenge and exploitation given any information differential. The level at which information is aggregated, and advantages derived, is an important consideration in determining the target of selection in an evolving system. The cohesion of groups, the cognitive ability of individuals to aggregate information, their positioning in social networks in order to acquire varied and timely knowledge (Bettencourt, 2003), and the recognition of temporary informational imbalances in markets are all examples of sources of individual and social advantage. These strategies of information search should clearly be taken into consideration in the study of decision processes and their social consequences.

The formalism of decision theory, although making reference to information, does not usually formalize it in the terms discussed here. Instead it builds maps between a space of decisions (or actions) and the utility (value) of each outcome. Game theory generalizes this structure to several players. The inclusion of information in these schemes necessarily enlarges decision processes to include components of learning, information flow and its exploitation. These processes may shed light on the mechanisms that bridge individual cognition and empirical work in behavioral sciences and the idealized optimal aggregate social behaviors central to economics and the social sciences. This synthesis is necessary both for the deeper understanding of social cognition and for its rigorous quantitative assessment as a force for change both in natural evolution and in human societies.

Acknowledgments

It is a pleasure to thank Robert Goldstone and Todd Gureckis for the invitation to contribute to this special volume. I thank them and four anonymous referees who helped make the manuscript better. Data on linguistic corpora has been harvested and processed by Michael S. Warren and Aric Hagberg at Los Alamos National Laboratory. The manuscript reflects parts of published work and several ongoing projects in collaboration with Aric Hagberg, Vadas Gintautas, Michel Ham, and Michael Warren. This work was partially supported by the Office of Science of the Department of Energy (USA) and the Los Alamos National laboratory LDRD program.

References

Ariely, D. (2009). *Predictably irrational: The hidden forces that shape our decisions*. New York: HarperCollins.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*, 1390.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis (Exposition of a New Theory on the Measurement of Risk). *Econometrica*, *22*, 23–36.

Bettencourt, L. M. A. (2003). cond-mat/0304321.

Bettencourt, L. M. A., Gintautas, V., & Ham, M. I. (2008) *Physical review letters* 100, 238701, arxiv.org/abs/0712.2218.

Bettencourt, L. M. A., Stephens, G. J., Ham, M. I., & Gross, G. W. (2007). *Physical Review E*, 75: 021915.

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm intelligence: From natural to artificial systems*. Oxford, England: Oxford University Press.

Bourgault, F., Furukawa, T., & Durrant, H. E. (2003) *Whyte Proceedings of the 2003 IEEURSJ InU. Conference on Intelligent Robots and Systems*, Las Vegas, Nevada, October 2003, pp. 48–53.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1*. Philadelphia: Linguistic Data Consortium.

Brown, P. F., Mercer, R. L., Della Pietra, V. J., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*, 467–479.

Camerer, C. F., Loewenstein, G., & Rabin, M. (Eds.) (2003). *Advances in behavioral economics (The roundtable series in behavioral economics)*. Princeton, NJ: Princeton University Press.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Couzin, I. D. (2009). Collective cognition in animal groups. *Trends in Cognitive Sciences*, *13*(1), 36–43.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Dennis, S. (2004). An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences USA*, *101*, 5206–5213.

Dorigo, M., & Stützle, T. (2004). *Ant colony optimization*. Cambridge, MA: MIT Press.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, *25*, 383–417.

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Germann, T. C, Kadau, K., Longini, I. M. Jr, & Macken, C. A. (2006). Mitigation strategies for paudemic influenza in the United States. *Proceedings of the National Academy Sciences USA*, *103*, 5935–5940.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using Collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*, 61–70.

Hölldobler, B., & Wilson, E. O. (2008). *The superorganism: The beauty, elegance, and strangeness of insect societies*. New York: W.W. Norton.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Luce, R. D. (2003). Whatever happened to information theory in psychology? *Review of General Psychology*, *7*, 183–188.

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. New York: Wiley.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

McElree, B., Frisson, S., & Pickering, M. J. (2006). Deferred interpretations: Why starting Dickens is taxing but reading Dickens isn't. *Cognitive Science*, *30*, 113–124.

Miller, G. A. (1956). The magical number seven. *Psychological Review*, *63*, 81–97.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact and information gain. *Psychological Review*, *112*, 979–999.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*, 1560–1563.

Parsons, S., Gymtrasiewicz, P., & Wooldridge, M. (Eds.) (2002). *Game theory and decision theory in agent-based systems*. Heidelberg, Germany: Springer.

Pederson, P., Dudenhoeffer, D., Hartley, S., & Permann, M. (2006). *Critical infrastructure interdependency modeling: A survey of U.S. and international research*. Idaho National Laboratory (INL) report NL/EXT-06-11464. Available at: http://www.osti.gov/bridge/product.biblio.jsp?osti_id=911792.

Pinker, S. (1999). How the mind works. *Annals of the New York Academy of Sciences*, *882*(1), 119–127.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Riley, S. (2007). Large-scale spatial-transmission models of infectious disease. *Science*, *316*, 1298–1301.

Ross, D. (2007). *Economic theory and cognitive science: Microexplanation*. Cambridge, MA: MIT Press.

Schneidman, E., Still, S., Berry, Michael J. II, & Bialek, W. (2003). *Physical Review Letters*, 91, 238701.

Seeley, T. D., & Visscher, P K. (2004). Quorum sensing during nest-site selection by honey bee swarms. *Behavioral Ecology and Sociobiology*, *56*(6), 594–601.

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). *Proceedings of the fifth annual workshop on computational learning theory* (pp. 287–294). Pittsburgh, PA, ACM.

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal*, *30*, 50–64.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Chicago: University of Illinois Press.

Simon, H. (1957). *Models of man*. Hoboken, NJ: John Wiley & Sons.

Simon, H. (1982). *Models of bounded rationality*, *Vols. 1–3*. Cambridge, MA: MIT Press.

Simon, H. (1991). On computable numbers, with an application to the Entscheidungs problem. *Journal of Economic Perspectives*, *5*, 28.

Sulton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambrigde, MA: MIT Press.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Random House.

Tapscott, D., & Williams, A. D. (2008). *Wikinomics: How mass collaboration changes everything*. New York: Penguin Group Incorporated.

Terano, T., Deguchi, H., & Takadama, K. (2003). *Meeting the challenge of social problems via agent-based simulation*. Heidelberg, Germany: Springer.

Turing, A. M. (1936). *Proceedings of the London Mathematical Society*, *ser. 2*, 42.

Vergassola, M., Villermaux, E., & Shraiman, B. I. (2007). Infotaxis as a strategy for searching without gradients. *Nature*, *445*, 406–409.

de Waal, F. B. M., & Tyack, P. L. (Eds.) (2003). *Animal social complexity: Intelligence, culture and individualized societies*. Cambridge, MA: Harvard University Press.

Wiener, N. (1948). *Cybernetics, or control and communication in the animal and the machine*. New York: Wiley.