What is Information?

Jerzy Konorski*
Faculty of Electronics, Telecomm. & Informatics
Gdansk University of Technology
80-952 Gdansk
Poland
jekon@eti.pg.gda.pl

Wojciech Szpankowski†
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Dedicated to our colleague and friend Jorma Rissanen, the philosopher of information theory

Abstract

The notion of *information* has so far been quantified mostly in statistical terms, giving rise to Shannon's information theory and the principles of digital data transmission. Studies of systems involving complex, intelligent, and autonomous agents, not uncommon in contemporary science, call for a new look at the measures of information that place importance on context, semantics, structures, and rationality. In this essay we propose a framework for measuring information inspired by the *event-driven approach*. We then illustrate our definition with several examples ranging from distributed computer systems to biology and economics.

1 Introduction

In this essay we muse on the notion of *information*, hoping to capture some of its essential aspects and provoke a discussion. We point out the need for a new definition of information that might be applied in contemporary science and engineering ranging from biology to chemistry, economics, and physics. We shall proceed inductively, giving examples from which hopefully a formal framework will arise.

Advances in information technology, the abundance of information systems and services, the much-trumpeted advent of information society, or even the Information Age (recently embodied in the communities of Web 2.0), almost obscure the fact that the common buzzword – the *i*-word – remains undefined in its generality, though considerable collective effort was harnessed into its understanding (cf. [6, 8, 20, 21, 31, 36, 38]). Shannon wrote in [32]: "The word "information" has been given many different meanings ... it is likely that at least a number of these will prove sufficiently useful in certain applications and deserve further study and permanent recognition."

Shannon's successful theory of information defines *statistical* information that quantifies to what extent a recipient of data can reduce statistical uncertainty associated with its source by observing the output of a source-recipient channel. Shannon also argued in his 1948 paper: "These semantic aspects of communication are irrelevant to the engineering problem." The channel error rate, on the other hand, does matter: for example, with a 50% binary error rate, the amount of statistical information sent through a binary symmetric channel is zero. But it

*The work of this author is sponsored by the AFOSR Grant FA8655-08-1-3018 and the Ministry of Science and Higher Education, Poland, Grant PBZ-MNiSW-02/II/2007.

†This work was supported in part by the NSF Grants CCR-0208709, CCF-0513636, and DMS-0503742, NIH Grant R01 GM068959-01, and the AFOSR Grant FA8655-08-1-3018.

seems that the intuitive understanding of information cannot be formalized without bringing into the picture the *timing* of data (consider a train departure notice served a recipient after the stated departure time), *spatial* aspect of information (imagine the same notice arriving at a different location), the *objective* its recipient wants to achieve (consider the same notice served a recipient not going anywhere), and the knowledge of the recipient's internal rules of conduct, or *protocol* for short (consider a recipient at the output of a channel with a high bit error rate, whose protocol dictates that the channel be regarded as perfect, hence received data be used *bona fide*).

The *context* of data cannot be abstracted from, either. Even at a high error rate some information may be recovered from the context e.g., a math textbook transmitted over such a channel might still be recognized as such. This point becomes particularly valid in the realm of biosystems – most biological information depends on where it is retrieved e.g., its location within a cell, a piece of DNA or protein. This important aspect is not yet well understood or analyzed in information theory. Biology is above all about context, and so a periodic pattern, while containing less statistical information than a random sequence, may contain a lot more *biological* information. In fact, in a recent paper [11] the authors argue that a random string and an exactly duplicated string add nothing or almost nothing to a biological information content. On the other hand, any context-dependent information measure must take into account the relationship between a given string and other related strings.

So what is information? In this essay, following C. F. von Weiszsäcker, we first argue that that there is *no absolute meaning* of information. Then, using an event-driven approach, we propose a definition that encompasses two of Weiszsäcker's premises, namely that "Information is only that which produces information" (relativity) and "Information is only that which is understood" (rationality) [36]. We then present some examples illustrating new aspects of information within the framework that we adopt here. We conclude with remarks suggesting some future work and leading to more questions. As a matter of fact, we hope to put forward some educated questions as to the issues and tools that lie before researchers interested in information, rather than come up with definite answers.

A preliminary version of this essay was prepared for the October 2005 workshop *Information Beyond Shannon* at Orlando FL. We thank the participants of the workshop for lively and constructive comments, some of which have found their way into the present version.

2 Event-Driven Approach

An intuitive relationship between data (any sequence of interpretable symbols) and information is that data may or may not carry information. One may observe that *a piece of data carries information if it helps its recipient achieve some objective.* In fact, this observation, stated more or less explicitly, was the point of departure of early textbooks on information technology [28]. There has been little formal apparatus, however, to quantitatively account for all its facets. To generalize and add precision we observe that *a piece of data carries information if it can impact a recipient's objective, under a given protocol and within a given context.*

Thus information has a flavor of *relativity* and *rationality*: it derives from the recipient's knowledge (gathered from the context), capability (implied by its protocol), and the pursued

objective. Underlying the latter are also temporal and spatial aspects, for the usefulness of data may depend on the timing and location of its generation and reception.

We offer more examples to illustrate the role of protocol. Clearly, a speaker of Chinese (a more knowledgeable recipient) can make out a lot more of a textbook on VLSI circuit design written in that language than a non-speaker (a less knowledgeable recipient). However, the latter can by default regard some strings of symbols that do not look like an ethnic language as a blueprint of a VLSI circuit; hence, the protocol can make up for the lack of knowledge (if applied only to the drawings in the textbook) or bring about catastrophic results (if applied to the Chinese characters of the text body). Furthermore, a duplicate notice of a train departure time does not contribute to the objective of catching that train and therefore is of no informational value (the recipient already knows it), unless the recipient's protocol stipulates that at least one confirmation of the train departure time be received. Finally, in a secret sharing scheme, decryption keys separated in time and space seem to carry zero information until they are brought together into one location at the same time. Indeed, information carried by data is not only related to its context, but also to a recipient's protocol, the rule dictating how to handle received data.

Having said this, we still need a quantitative definition, an analogue of Shannon's statistical information, retaining the flavors of relativity and rationality, and with a potential to reflect temporal and spatial aspects. Can we attempt formal definitions of the *amount* of information and maximum amount of information carried by a channel – capacity – without a lengthy specification of the semantics of data? One possibility is to adopt an event-driven approach which we sketch below.

An event-driven approach offers a few advantages. First, it is well-established among the engineering community thanks to the work of C. A. R. Hoare and others in the field of operating systems and distributed algorithms. Second, it is discrete and timeless in nature, yet allows for dynamic characterization of systems evolving in continuous time. Finally, it is able to formalize such intuitions as causality and consistency of local views without specifying the semantics of the involved events. At the same time, it generalizes the data-information relationship: now it is events that may or may not carry information; in particular, an event may correspond to reception of a piece of data, a clock tick etc. The event-driven approach-inspired formalization goes along the following lines:

- A universe is populated by systems (living organisms, institutions, communities, software agents, Internet domains etc.) pursuing specified *objectives*.
- A system's current state is expressible through a number of system variables (e.g., memory content, parameter configuration, operational status of constituent subsystems); an observable change of state marks an *event* (e.g., clock tick, execution of a specific operation, reception of a piece of data from another system).
- A partial order on the set of events may be defined as the order in which the events occur at a given system (with simultaneous events not precluded); the set of events preceding an event is called the *context* of the event.
- Events may have attributes e.g., time of occurrence and semantics, as defined by the

system's *protocol* i.e., specification of how the system handles the events in order to pursue its objectives.

We would like to regard information as another (measurable) attribute of an event reflecting our previous discussion. To this end, define an objective functional that maps a system's protocol P and a context C (a sequence of events related to the communication between the source and recipient systems) into any space with ordered points; further we only consider the one-dimensional Euclidean space i.e., real axis. The idea is that P along with C determine $objective(P, C)$, the extent to which the recipient system's objective has been achieved. For simplicity assume that P remains fixed throughout the system's lifetime. In particular, monotonicity of $objective(P, C)$ in C is desirable, for it implies that successive events help achieve the objective. That is, we would like $objective(P, C + E) \geq objective(P, C)$ for any event E and context C, where $C + E$ is the new context extended by event E. Before defining a possible measure of information we discuss more examples to support our approach.

Example 1. [*Decimal Representation*] Assume that a system's objective is to learn the number π and P has the system compute successive decimal digits approximating π from below. Each computed digit is then regarded as an event and $objective(P, C)$ is a real-valued function monotonically increasing and asymptotically stabilizing in C. As an illustration, imagine we are drawing circles of circumferences $3, 3.1, 3.14, 3.141$ etc., and measure the respective diameters i.e., $.9549, .9868, .9995, .9998$, which asymptote to the ideal 1.

Example 2. [*Shannon Information*] In Shannon's information theory [31] *objective* is defined as *statistical ignorance of the recipient* or *statistical uncertainty of the recipient*. It is measured by the number of binary decisions to recognize the event E, that is, $-\log P(E)$, where $P(E)$[1] is the probability as computed by the recipient. For various generalizations the reader is referred to [18, 21]. Observe also that spatial and temporal aspects of information were mostly left out in Shannon's theory.

Example 3. [*Distributed Information*] In an (N, N)-threshold secret sharing scheme [29], N subkeys of the decryption key roam among geographically dispersed systems. By the protocol P, the event corresponding to the reception of another subkey from a fellow system does not give access to the secret unless receptions of *all* the other subkeys are already in C. Likewise, an observed pixel of a digital image may increase a viewer's ability to understand the image depending on how many neighboring pixels have already been observed (this example illustrates that the event-driven approach also covers spatial, rather than temporal, contexts – in general, there is no difficulty evaluating the objective functional as long as events are processed sequentially). In passing we may wonder what is the difference between distributed and local information; is one bit *here* equivalent of one bit *there*?

Example 4. [*Temporal Information*] The impulses exchanged along nerves or processed within neural cells of a living organism critically depend on timing e.g., a stimulus generated by a pain

[1] We shall write $P(E)$ for the probability of an event E since from the context one easily distinguishes it from the protocol P.

receptor is useless if it arrives too late to administer a defensive gesture. Spatio-temporal coding is widely acknowledged to be the most important information processing feature of networks of neurons [17]. This remarkable coding scheme forces groups of neurons, involved in the same learning or memory retrieval task, to communicate and process information through *timing* and *location*. The spatial aspect of this form of coding arises due to functional differentiation of neurons. Usually, neurons involved in processing of related tasks or designed to respond to similar cues are clustered in the same region of the brain. Examples include the well known receptor maps in the olfactory bulb, the cochleotopic (frequency) regions in the primary auditory cortex (where different regions of neurons respond to different frequencies in the stimulus), and the topographic feature maps in the visual area of mammalian brains (where neurons discriminate against different orientations of the visual stimulus).

Similarly, clock ticks are relevant when judging the usefulness of successive speech or video frames sent over a packet network. Since they share network resources with unpredictable data traffic, the frames arrive at the destination irregularly, as quantified by delay jitter. Premature and overdue arrivals (events with too few or too many clock ticks in the context) are equally unwelcome, though are handled in a different way: the former have to be buffered before delivery and the latter are typically discarded. In general, incurred delay (e.g., in biological and computer networks) is a nontrivial issue not yet successfully addressed by information theory [14].

Example 5. [*Wireless Networks*] In a wireless ad hoc network, each mobile terminal (MT) can physically communicate only within its transmission range. To maintain network-wide connectivity and so achieve the objective of each MT (i.e., a high throughput of data packets), P prescribes setup and maintenance of relay paths between remote MTs. These are temporary in nature due to the terminal mobility. Thus there are both path discovery and path disruption events; consequently and somewhat counterintuitively, $objective(P, C)$ may not increase in C. Recent research [12, 13, 15] indicates that for $objective(P, C)$ to increase in C, a quite unorthodox P is needed that restricts paths to two-hop and trades buffer space for bandwidth, a thought at the core of the so-called time capacity paradox.

Example 6. [*Herding, Web 2.0, DNA*] The conclusion of the previous example suggests that $objective(P, C)$ increases in C provided that P is somehow "rational." Unfortunately, studies of the so called *herding effects* disprove that intuition too: an individual contemplating an action behaves rationally by observing and following the majority of other individuals (as shown by Bayesian analysis). After a short while, however, further observations provide no more insight into the benefits of the action [4]. Perhaps, then, one can only assert that $objective(P, C)$ is nondecreasing in C provided that P is rational? There are examples that run counter even that intuition. Imagine a user session with a Web search engine in which too much data, or the presence of conflicting data, paralyze the user's ability to act; from another perspective, a growing number of users contributing their ideas to a digital Web 2.0 community may at some point prevent a required broad consensus. Equally daunting is the well-known fact that the sheer amount of data contained in a biological database (e.g., human genome) may blur patterns leading to the identification of relevant human traits. In fact, in a massive data set, such as a biological database or results of an Internet search, the situation is not unlike a radio channel crossed by interfering signal paths: what is noise for one receiver (query) may well be

useful information for another. The problem of discovering and quantifying the amount of useful information thus acquires a new meaning.

Example 7. [*Cooperative and Noncooperative Settings*] Consider now a system where the objective functionals defined at different subsystems are in conflict (e.g., the problem of Byzantine generals, DoS or selfish attacks on communication protocols such as IEEE 802.11 [19]). The simplest example are two data sources contending for a multiple access channel (e.g., ALOHA system). Various forms of P may then calibrate the sources' behavior from cooperative (where *objective*(P, C) increases in the total number of data transmission events in C i.e., in the overall channel utilization) to noncooperative (where *objective*(P, C) increases in the number of own data transmission events) to malicious (where *objective*(P, C) decreases in the number of the other source's data transmission events).

Example 8. [*Rissanen's Stochastic Complexity and MDL*] Included in *objective*(P, C) may be the cost of the very recognition and interpretation of C. Imagine a recipient knowing that the source uses an optimal code for its stream of data, but having to learn on the fly the stochastic mechanism according to which the source generates data. As time passes, the model reveals itself to the recipient who can then hypothesize about data sent. In 1978 Rissanen [23, 24, 25, 27] introduced the *Minimum Description Length* (MDL) principle, an incarnation of Occam's Razor stating that the best hypothesis is the one that gives the shortest description of data. Realizing that Kolmogorov complexity is uncomputable, MDL selects a code for which the *total* description length of code and data is minimal. Rissanen stresses that we should "make no assumptions" about a *true* data generating process. In practice, we must restrict the class of process models.

More precisely, let $\mathcal{M}_k = \{Q_\theta : \theta \in \Theta\}$ be a set of finitely parameterized distributions of dimension k. One could argue, and some did, that the best (shortest) description of a string $x = (x_1, \ldots, x_n)$ should be $-\log Q_\theta(x)$, as suggested by the Kraft correspondence for prefix codes. As pointed out by Rissanen and others, this is not correct since one must also describe the distribution Q_θ itself. But this can be accomplished by a *universal data compression* algorithm. Rissanen proposed two possible solutions, namely *two-part codes* and the *normalized maximum likelihood* (NML) code that we briefly describe below.

In the two-part coding, one first describes a distribution Q_θ and then describes the string x using Q_θ. Let $\mathcal{C}$ be a code that maps Θ to $\{0,1\}^*$. Then the *stochastic complexity* $S(x)$ is

$$S(x) = \min_{\theta \in \Theta} \left[\text{length}(\mathcal{C}(\theta)) - \log Q_\theta(x) \right],$$

and the MDL principle states that one should choose θ^* that achieves the above minimum.

In the normalized maximum likelihood (NML) code, first the parameter $\hat{\theta}$ is chosen to minimize $-\log Q_\theta(x)$ (as in the classical maximum likelihood estimate), and then the "ideal" code-length $-\log Q_{\hat{\theta}}(x)$ is used as a yardstick against which code performance is measured. This leads to the so called *minimax problem* that finds the best code for the worst distribution and the worst data. It is well known [3, 9, 26] that the regret function defined as

$$r_n^*(\mathcal{M}) = \min_Q \max_x \left[\log \frac{Q_{\hat{\theta}}(x)}{Q_\theta(x)} \right]$$

achieves its optimal value $\log \sum_x Q_{\hat{\theta}}(x)$ for the *normalized maximum distribution*

$$Q_{NML}(x) = \frac{Q_{\hat{\theta}}(x)}{\sum_x Q_{\hat{\theta}}(x)}.$$

The optimal code-length is then $-\log Q_{NML}(x)$. Rissanen in [26] proved, among others, that the minimax regret for $\mathcal{M}_k$ is

$$r_n^*(\mathcal{M}_k) = \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int_\theta \sqrt{|I(\theta)|} d\theta + o(1)$$

where $I(\theta)$ is the Fisher information. Further generalization can be found in [3, 9, 16, 33]. In passing, one still may ask why to restrict analysis to prefix codes? Is there a fundamental lower bound for general codes (cf. [1, 34, 37])?

3 Information and Capacity

We are now in a position to set out a framework for defining the amount of information consistent with the intuition based on our examples and discussion.

Definition 1 *The amount of information carried by event E in context C as perceived at a system with protocol P is*

$$\text{info}_{P,C}(E) = \text{weight}[\mathit{objective}(P, C + E), \mathit{objective}(P, C)], \tag{1}$$

where "weight" measures the change between two (objective) points according to the order defined on the space of values of the objective functional.

Thus an event only carries nonzero information if it changes $\mathit{objective}(P, C)$, a statement consistent with the intuitive flavors of relativity and rationality. The dependence on P and C reflects the obvious observation that one and the same event can produce different information at different recipients, locations, and times. Also note that in view of Example 6, negative information is not unthinkable. In fact, this might lead to an interesting distinction: *nonconfoundable* systems, contrasted with *confoundable* ones, are those whose protocol P precludes negative information regardless of C. One can imagine a smart Web user always able to remove conflicting data from the context and proceed monotonically towards an objective. Whether and for what types of data sources and objective functionals such P exist is an open problem. Finally, it is natural to surmise that both P and C are subject to various constraints implied, respectively, by the systems' architecture and the nature of the event sources. In the spirit of Shannon, one may define the channel capacity between the event source and the recipient as a maximum-type measure on a collection of amounts of information carried by successive events, within the regions of feasible P and C (subject to the said constraints). For a given $C = (E_1, E_2, \ldots)$ and $E_i \in C$, let $C_i := (E_1, \ldots, E_{i-1})$ be the prefix of C consisting of events preceding E_i.

Definition 2 *The capacity of the channel between the event source and recipient is*

$$\text{capacity} = \max_{P \text{ feasible}} \max_{C \text{ feasible}} F\left(\{\text{info}_{P,C_i}(E_i),\ i \geq 1\}\right). \tag{2}$$

for some function $F(\cdot)$.

Depending on the specific case, the function F can be conveniently defined as the sum of all elements of its set argument, the maximum element, etc. If the total amount of information and the feasible C are infinite, it may be convenient to define F as the limiting average information per event:

$$F\left(\{\text{info}_{P,C_i}(E_i),\ i \geq 1\}\right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \text{info}_{P,C_i}(E_i) \tag{3}$$

provided the limit exists. With so structured a definition it is possible to confine interest to the inner maximum if for some reasons P is regarded as the only feasible.

We now return to some of the previous examples in order to give a quantitative illustration of Definitions 1 and 2.

Example 1. [continuation] In Example 1, the objective in a given context can be measured as the deviation of the corresponding diameter from the ideal 1, so that the amount of information carried by successively computed digits of π is the difference between successive deviations. Hence, the event "3" carries $(1-0)-(1-.9549) = .9549$, "1" carries $(1-.9549)-(1-.9868) = .0319$, "4" carries $(1-.9995)-(1-.9868) = .0127$, the other "1" carries $(1-.9998)-(1-.9995) = .0003$ units of information etc. If F is as in (3), then the capacity of such a channel is zero: an infinite number of events carry a finite total information.

Example 2. [continuation: *Shannon Information and Temporal Capacity*] Does the event-driven approach include Shannon information as a special case? As suggested by the previous discussion, the objective in Shannon information can be viewed as the negative of statistical uncertainty. Consider a memoryless channel and a memoryless source transmitting symbols chosen from a finite set according to some probability distribution. The amount of information carried by an event $E = (x, y)$, where x and y are respectively the transmitted and received symbol, can be measured by the difference between the recipient's degree of certainty as to x before and after reception of y i.e.,

$$\text{info}_{P,C}(E) = -\log P(x) - [-\log P(x|y)].$$

Note that because of our memoryless setting, there is no explicit dependence on C. If the channel is noiseless (error-free), then $P(x|y) = 1$ iff $x = y$, thus $\text{info}_{P,C}(E) = -\log P(x)$. Taking F in our definition of capacity as in (3), we find for a context $C = (E_1, \ldots, E_n)$

$$F\left(\{\text{info}_{P,C_i}(E_i),\ i \geq 1\}\right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \text{info}_{P,C_i}(E_i) = -\sum_{x} P(x) \log P(x) = H(X).$$

Here, X is a random variable describing the source. The right-hand side of the above relationship we recognize as Shannon's entropy of the source. In a noisy channel, the limiting average information per event becomes

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \text{info}_{P,C_i}(E_i) = \sum_{(x,y)} P(x,y)[\log P(x|y) - \log P(x)] = I(X;Y),$$

where Y is a random variable describing the output of the channel. This we recognize as Shannon's mutual information. It is easy to see now that, with the protocol P fixed and the

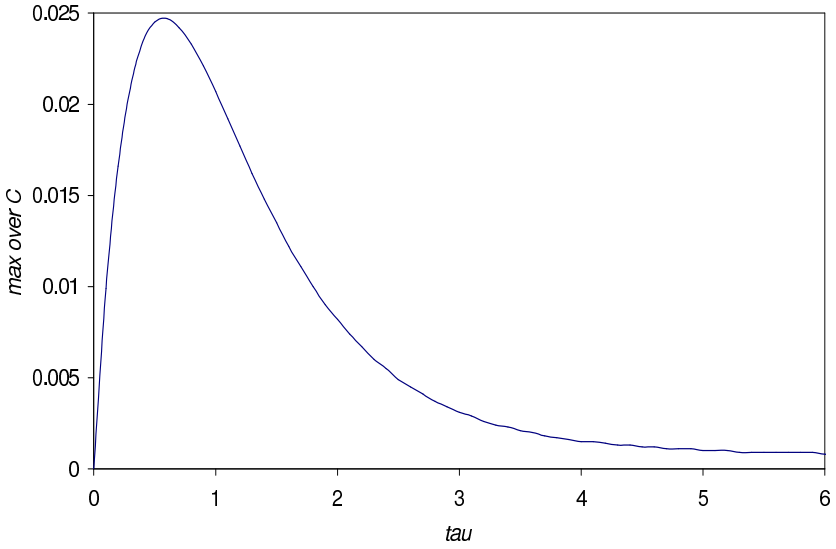

Figure 1: Temporal capacity as a function of τ.

maximization only taken over C, the channel capacity in the sense of Definition 2 coincides with Shannon's capacity

$$\max_{C \text{ feasible}} \frac{1}{n} \sum_{i=1}^{n} \text{info}_{P,C_i}(E_i) \sim \max_{P(X)} I(X;Y),$$

where the right-hand side maximum is taken over all possible distributions of X. This is so because in our memoryless setting, any feasible context must have been produced by some $P(X)$.

Recall that Shannon's celebrated channel coding theorem states that as long as the transmission rate does not exceed the channel capacity, information can be sent with as small a frequency of errors as desired provided unlimited time and resources are available to encode and decode the message. Thus, temporal (or spatial) aspects of information are not considered. However, they can easily be addressed in this setting, and the relevance of optimizing the protocol can be demonstrated.

Consider a memoryless binary symmetric channel with "temporal errors": the longer a binary symbol takes to reach the recipient, the lower the probability of a successful transmission. Each transmitted symbol is received in error with probability $\Phi(\varepsilon, t)$, where ε is the "instant error" rate and t is the incurred channel delay. A plausible function Φ should increase from 0 to 1 for $\varepsilon \in (0,1)$, and increase from ε to 1 as t varies between 0 and ∞. Assume further that the recipient's protocol P enables determination of t when a symbol is received, and if $t \geq \tau$ prescribes erasure of the received symbol. Thus $X \in \{0,1\}$ and $Y \in \{0,1,erasure\}$. Let the source be memoryless with $P(X=1) = p$ and the channel delay be represented by a random variable D with a known probability distribution function $F(t) = P(D < t)$. We only need to slightly modify the amount of information carried by an event $E = (x,y)$, namely

$$\text{info}_{P,C}(x,y) = \begin{cases} \log P(x|y,\ D < \tau) - \log P(x) & \text{if } y = 0,1 \\ 0 & \text{if } y = \text{erasure.} \end{cases}$$

Then the limiting average information per event again coincides with the mutual information

$I(X;Y)$. To calculate the latter let us introduce the conditional probability

$$\phi := P(Y=1|X=0, D<\tau) = P(Y=0|X=1, D<\tau) = \frac{\int_0^\tau \Phi(\varepsilon,t)dF(t)}{F(\tau)},$$

which plays the role of the "temporal error" rate. Standard calculation yields

$$I(X;Y) = [H_b((1-\phi)\cdot(1-p)+\phi\cdot p) - H_b(\phi)]\cdot F(\tau), \tag{4}$$

where $H_b(u) = -u\log u - (1-u)\log(1-u)$ is the binary entropy function for $u \in [0,1]$. By a similar argument as above, the maximization of (4) over p corresponds to maximization over feasible C in (2). The maximum is attained at $p = 1/2$ and yields

$$[1 - H_b(\phi)]F(\tau),$$

the maximum mutual information for a given τ, that is, Shannon channel capacity. In Figure 1 we plot this quantity against τ assuming $F(t) = 1-e^{-t}$ and $\Phi(\varepsilon,t) = 1-(1-\varepsilon)^{t+1}$ for $\varepsilon = 0.3$. (Note that the average channel delay is the time unit.) We see that in the case of a stringent delay bound the capacity of the channel is adversely affected by frequent erasures; when the delay bound becomes ineffective, frequent temporal errors dominate infrequent erasures to produce a somewhat counterintuitive drop in mutual information. We now recall that τ represents the recipient's protocol P; hence if we maximize over τ, which corresponds to the outer maximum in (2), we get a clear estimate of the channel capacity.

Example 3. [continuation] Let N subkeys move at random and independently of one another among $A \times A$ stations regularly spaced within a square area. For simplicity let the movements of

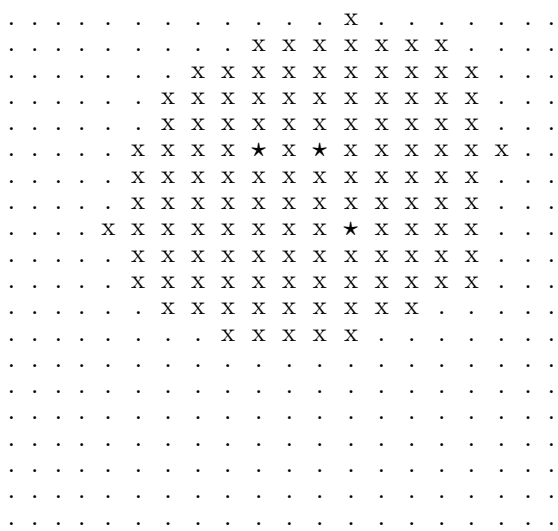

Figure 2: Access to the secret ($N = 3, A = 20, d = 8$).

the subkeys be synchronized to unit time slots. In each slot a station can improve its objective by having temporary access to the secret, which happens if it is within Euclidean distance d from each of the subkeys. This is illustrated in Fig. 2, where the current subkey locations are marked "*" and stations with access to the secret are marked "x."

Assume that the larger N, the more valuable the secret, which results in each "x" station improving its objective proportionally to N. If all the stations act as one system, then an event E defines N new locations of the subkeys. Here, N is a parameter of the set of feasible C, d is a parameter of P, and

$$\text{info}_{P,C}(E) = N \times \{\#\text{ of stations having access to secret}\}.$$

Note that since the movement of the subkeys is memoryless, there is no explicit dependence on C. The limiting average information per event per station, which thus equals N times the probability of access per event, is plotted in Figure 3 (obtained by a Monte Carlo simulation). The maximum of each curve corresponds to the channel capacity as expressed by the inner maximum in (2) i.e., with respect to C, given P.

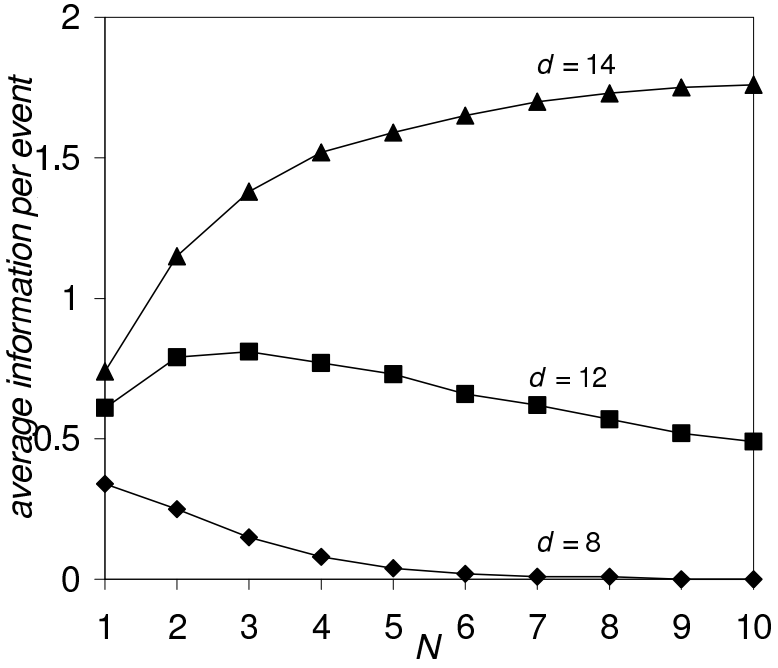

Figure 3: Normalized average information per event for secret sharing ($A = 20$).

Example 7. [continuation: *Noncooperative Settings; Value of Information*] We should point out that calculating the capacity in the above framework seems to be particularly difficult in a distributed system featuring multiple autonomous agents. For example, in economics one often considers the value of information [21] which measures (perhaps in dollars) the difference between the payoffs of an informed action and an uninformed action. Consider a simple entry deterrence game [10]. Suppose an Internet service provider (ISP) has a major business client (Incumbent) who can use either Standard or Premium service. Another business (Entrant) is considering entry i.e., becoming the ISP's client with only Standard service available. Both Incumbent and Entrant choose their strategies (Standard/Premium and enter/not enter) simultaneously and without prior coordination. Thus a one-shot noncooperative game arises with payoffs given in Table 1. Here, K is the surcharge Incumbent pays for Premium service ($0 \leq K \leq 3$). While the other payoff components are rather arbitrary, the relationships between them are important:

Table 1: Entry deterrence payoffs (arbitrary units): Incumbent's (*left*) and Entrant's (*right*)

	enter	not enter
Premium	$3-K,\quad -1$	$5-K,\quad 0$
Standard	$2,\quad 1$	$3,\quad 0$

- Entrant's payoff is neutral if she does not enter, otherwise it is negative if Incumbent chooses Premium (Entrant pays entrance fee, but receives a less-than-fair share of ISP's resources), and is positive if Incumbent chooses Standard (Entrant receives a fair share of ISP's resources),
- Incumbent is better off if Entrant does not enter (there is no competition for ISP's resources), and given Entrant's choice, Incumbent's well-being depends on K e.g., $K > 2$ ($K < 1$) makes Standard (Premium) a dominating strategy.

It is easy to see that for $K > 2$ the only Nash equilibrium (NE) is (Standard, enter), while for $K < 1$ the only NE is (Premium, not enter). For $1 \leq K \leq 2$ there exists a unique NE in mixed strategies:

$$\left(\frac{1}{2} * \text{Premium} + \frac{1}{2} * \text{Standard}, (1-(K-1)) * \text{enter} + (K-1) * \text{not\_enter}\right),$$

where $p * s + q * s'$ denotes a mixed strategy "play s with probability p and s' with probability q."

A more realistic model assumes that (a) Entrant has only an estimate K' of K, (b) Incumbent knows both K and K', moreover, is in a position to communicate K to Entrant if she thinks it worthwhile. The question is whether and when Incumbent will indeed communicate K and how much information passes between Incumbent and Entrant. Let E_0^{Ent} and E_0^{Inc} denote the events of acquiring the knowledge of K' by Entrant, and of K and K' by Incumbent. The objective is the expected payoff and the protocols of both players prescribe NE strategies. For Entrant, the NE strategy is:

$$s(E_0^{Ent}) = \begin{cases} \text{not enter}, & K' < 1 \\ [1-(K'-1)] * \text{ enter} + (K'-1) * \text{ not\_enter} & 1 \leq K' \leq 2 \\ \text{enter} & K' > 2. \end{cases}$$

Incumbent, who knows the above strategy, chooses hers so as to maximize the expected payoff. The result is obvious except when $1 \leq K \leq 2$ and $1 \leq K' \leq 2$. Incumbent's expected payoff conditioned on choosing Premium is then

$$[1-(K'-1)] \cdot (3-K) + (K'-1) \cdot (5-K)$$

and conditioned on choosing Standard is

$$[1-(K'-1)] \cdot 2 + (K'-1) \cdot 3.$$

Incumbent chooses Premium if the former payoff is greater than the latter, i.e., if $K < K'$, and Standard if $K > K'$. (If $K = K'$, Incumbent plays 1/2 * Premium + 1/2 * Standard.) Hence,

Table 2: Expected payoffs at NE: Incumbent's (*left*) and Entrant's (*right*)

	$K' < 1$	$1 \le K' \le 2$	$K' > 2$
$K > 2$	3, 0	$[1-(K'-1)]\cdot 2 + (K-1)\cdot 3$, $[1-(K'-1)]\cdot 1$	2, 1
$1 \le K \le 2$	$5-K$, 0	**as above** (if $K > K'$) $[1-(K'-1)]\cdot 2 + (K'-1)\cdot 3$, 0 (if $K = K'$) **as below** (if $K < K'$)	2, 1
$K < 1$	$5-K$, 0	$[1-(K'-1)]\cdot(3-K) + (K'-1)\cdot(5-K)$, $[1-(K'-1)]\cdot(-1)$	$3-K$, -1

$$s(E_0^{Inc}) = \begin{cases} \text{Premium} & K < 1 \text{ or } (1 \le K \le 2 \text{ and } K' < 1) \\ & \text{[Entrant does not enter]} \\ & \text{or } (1 \le K \le 2 \text{ and } 1 \le K' \le 2 \text{ and } K < K') \\ & \text{[Entrant plays mixed strategy]}, \\ 1/2 * \text{Premium} + 1/2 * \text{Standard} & 1 \le K \le 2 \text{ and } 1 \le K' \le 2 \text{ and } K = K', \\ \text{Standard} & K > 2 \text{ or } (1 \le K \le 2 \text{ and } K' > 2) \\ & \text{[Entrant enters]} \\ & \text{or } (1 \le K \le 2 \text{ and } 1 \le K' \le 2 \text{ and } K > K') \\ & \text{[Entrant plays mixed strategy]}. \end{cases}$$

For the payoffs in Table 1, the possible Incumbent's and Entrant's expected payoffs at NE are given in Table 2, which both players can compute using game theory basics, but only Incumbent knows which row gives actual payoffs.

Imagine now that just before the game, Incumbent has a chance to communicate K and thus correct Entrant's wrong estimate K' (denote the corresponding event E_1^{Ent}). This she will not consider worthwhile if $K > 2$ and $K' \le 2$ for it would encourage Entrant's entry, thereby decreasing Incumbent's expected payoff (from 3, or a value between 2 and 3, to 2). Similarly for $1 \le K \le 2$ and $K' < 1$. If $1 \le K \le 2$ and $K' > 2$, the communication of K would lead to the mixed strategy NE; this will increase Incumbent's payoff (from 2 to a value between 2 and 3), but at the same time decrease Entrant's payoff (from 1 to 0). If Entrant is noncooperative, she will ignore E_1^{Ent} regarding it as incredible (presumably part of Incumbent's entry deterrence strategy). Knowing that, Incumbent will simply communicate nothing. In all the above cases, the channel between Incumbent and Entrant is as good as closed (unable to carry information).

Only when $K < 1$ and $K' \ge 1$ will the communication of K become worthwhile from Incumbent's viewpoint and credible to Entrant, for Entrant's expected payoff then would rise (from -1, or a value between -1 and 0, to 0). According to (1), the amount of information received by Entrant in this case is:

$$\begin{aligned}\text{info}(E_0, E_1) &= \text{payoff}(E_0^{Ent}, E_1^{Ent}) - \text{payoff}(E_0^{Ent}) \\ &= \begin{cases} 0-(-1)=1, & K' > 2 \\ 0-(-1)\cdot[1-(K'-1)] = 1-(K'-1), & 1 \le K' \le 2. \end{cases}\end{aligned}$$

If the game is played repeatedly, e.g., on a session basis, and each time K and K' are drawn independently from a uniform probability density function on $[0, 3]$, then their joint probability density is $1/9$. The average amount of information received by Entrant (i.e., the average increase in Entrant's objective) per game is:

$$\text{avg\_info} = \frac{\int_0^1 \left[\int_1^2 [1-(K'-1)]dK' + \int_2^3 1dK'\right] dK}{9} = \frac{3}{2\cdot 9}$$

What if neither Incumbent nor Entrant were noncooperative and so K were communicated and E_1^{Ent} were accepted regardless of the expected payoffs? Then

$$\begin{aligned}9\cdot\text{avg\_info} &= \int_0^1 \left[\int_1^2 [1-(K'-1]dK' + \int_2^3 1dK'\right] dk + \int_1^2 \left[\int_0^1 0dK' + \int_2^3 (-1)dK'\right] dK \\ &\quad + \int_2^3 \Bigg] \int_0^1 1dK' + \int_1^2 (K'-1)dK'\Bigg] dK = 2,\end{aligned}$$

thus avg_info= $2/9$. In summary, the 25% difference between the latter two figures reflects the reduction of channel capacity merely due to noncooperative nature of the involved protocols

4 Final Remarks

Our definition (1) is somewhat similar in spirit to that of the *value of information* discussed in Luenberger [21]. In the presence of a single source of uncertainty about the state of the world among the many possible states, Luenberger considers a decision-maker maximizing the average payoff and calculates the net benefit of receiving an imperfect signal about the true state of the world. Clearly, the net benefit is zero if the signal does not reduce the uncertainty. In such a Bayesian setting, negative values of information are impossible. Although we propose a broader framework, with context and protocol explicitly accounted for, we still need a generalization of imperfect signals (or imperfect events in our wording); ours is a faultless communication system, where events do not get corrupted or misinterpreted. While partly justified by contemporary high-quality transmission and processing infrastructure, this is a serious restriction.

A no less fundamental issue is related to the very notion of information. The foregoing discussion focused upon *communicable* information, which is why events played so central a role: a system remaining in one and the same state cannot change its perception of the achieved objective. However, another strong intuition of information holds it to be embedded in the structure of an object and thus independent of any rational activity – this we may refer to as *structural information*. F. Brooks articulates in [5] : "Shannon and Weaver performed an inestimable service by giving us a definition of information and a metric for information as communicated from place to place. We have no theory however that gives us a metric for

Figure 4: Access to the secret with two different subkey locations ($N = 3, A = 20, d = 8$)

the information embodied in structure ... this is the most fundamental gap in the theoretical underpinning of information and computer science. ... A young information theory scholar willing to spend years on a deeply fundamental problem need look no further." Along with spatial and temporal aspects of information, this is, in our opinion, the most urgent challenge facing our community.

Yet another understanding arises from a conjecture of an organizing principle, a hidden mechanism behind a given object, and the amount of structural information may be related to the remaining uncertainty as to the nature or parameters of the hidden mechanism. In this way a sequence with clear patterns of symbols may be attributed more structural information than a piece of gibberish after all. This is particularly true about biological information as discussed above and in [11]. To illustrate our point, consider again our secret sharing scheme, as in Example 3, and suppose we only know the current "x" stations. Two sets of such stations, corresponding to two different subkey locations, are depicted in Fig. 4. They may be regarded as two states of our system, or two objects of some informational value, the hidden mechanism being the movement of the subkeys. Where can the subkeys be? They can be no further than d from any "x" station, which leaves a number of possible subkey locations marked "?" in Fig. 5. (Particular N-tuples of locations can then be eliminated at the cost of more computation.) We might conclude that the left state (object) contains more structural information than the right one.

In summary, we propose to fundamentally enhance six decades of work in information theory by incorporating the following elements that were, to large extent, not adequately addressed in the past and therefore threaten to raise severe impediments to diverse applications:

Structure: We still lack measures and meters to appraise the amount of organization and information embodied in artifacts and natural objects.

Delay: In typical interacting systems, timeliness of signals is essential to function. Often timely delivery of partial information carries higher value than delayed delivery of complete information. For example, in a signaling cascade associated with a specific cell function, delay or loss of signals

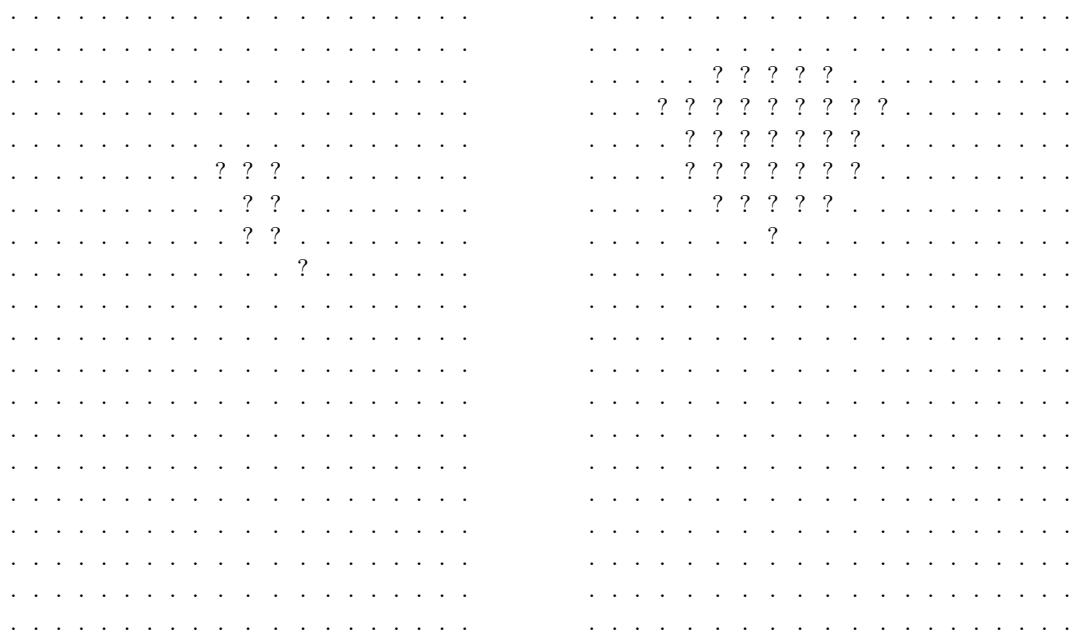

Figure 5: Possible subkey locations

can be lethal.

Space: In interacting systems, spatial localization often limits information exchange – with obvious disadvantages as well as benefits. These benefits typically result from reduction in interference (common examples range from wireless systems to immune response).

Information and control: In addition to delay-bandwidth tradeoffs discussed above, systems often allow modifications to underlying design patterns (e.g., network topology, power distribution and routing in networks). Simply stated, information is exchanged in space and time for decision making, thus timeliness of information delivery along with reliability and complexity constitute basic objectives.

Semantics. In many scientific contexts, one is interested in signals, without knowing precisely what these signals represent (e.g., DNA sequences, spike trains between neurons, whale songs), but little more than that can be assumed a priori. Is there a general way to account for the actual "meaning" of signals in a given context?

Dynamic information. In a complex network, information is not just communicated but also processed and even generated along the way. How can such considerations of dynamic sources be incorporated into an information-theoretic model?

Learnable information. One may argue (and some have) that in all scientific endeavors, the only task is to extract information from data. How much information can actually be extracted from a given data repository? In Shannon theory, one starts from a (possibly unknown) model for the data-generating mechanism and calculates its entropy, but in practice the starting point is only the data. Is there a general theory that provides natural model classes for the data at hand? What is the cost of learning the model, and how does it compare to the cost of actually describing the data?

Limited Resources: In many scenarios, information is limited by available resources (e.g., computing devices, living cell). How much information can be extracted and processed with limited resources?
Quantum Information: Microscopic systems do not seem to obey Shannon's postulates of information. In the quantum world and on the level of living cells, traditional information often fails to accurately describe reality [6].
Value of Information: The impact of rational and noncooperative behavior upon information as well as the value of information, should be studied in more generality.

References

[1] N. Alon and A. Orlitsky, A Lower Bound on the Expected Length of One-to-One Codes, *IEEE Trans. Information Theory*, 40, 1670-1672, 1994.

[2] H. V, von Baeyer, *Information: The New Language of Science*, Harvard University Press, 2004.

[3] A. Barron, J. Rissanen, and B. Yu, The Minimum Description Length Principle in Coding and Modeling, *IEEE Trans. Information Theory*, 44, 2743-2760, 1998.

[4] S. Bikhchandani, D. Hirshleifer, I. Welch, Learning from Others: Conformity, Fads, and Informational Cascades, *J. Econ.Perspectives*, 12 151-170, 1998.

[5] F. Brooks, Three Great Challenges for half-century-old computer science, *J. the ACM*, 50, 25-26, 2003.

[6] C. Brukner, A. Zeilinger, Conceptual Inadequacy of the Shannon Information in Quantum Measurements. *Phys. Rev.* A 63, 2001.

[7] C. Cherry, *On Human Communication*, The MIT Press, Cambridge, Massachusetts, 1978.

[8] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Second Edition, John Wiley & Sons, New York, 2006.

[9] M. Drmota and W. Szpankowski, Precise Minimax Redundancy and Regret, *IEEE Trans. Information Theory*, 50, 2686-2707, 2004.

[10] D. Fudenberg and J. Tirole, *Game Theory*, Cambridge, MA: MIT Press, 1991.

[11] D. Galas, M. Nykter, G. Carter, N. Price and I. Shmulevich, Set-based Complexity and Biological Information, preprint 2007.

[12] M. Grossglauser and D. Tse, Mobility Increases the Capacity of ad-hoc Wireless Networks, *IEEE/ACM Trans. Networking*, 48, 477-486, 2002.

[13] P. Gupta and P.R. Kumar, Capacity of Wireless Networks, *IEEE Trans. Information Theory*, 46, 388-404, 2000.

[14] B. Hajek and A. Ephremides, Information Theory and Communication Networks: An Unconsummated Union, *IEEE Trans. Information Theory*, 44, 2416-2434, 1998.

[15] P. Jacquet, Space-time Information Propagation in Mobile ad hoc Wireless Networks, `http://ee-wcl.tamu.edu/itw2004/program/jacquet_inv.pdf/`, *ITW 2004*, San Antonio, 2004

[16] P. Jacquet and W. Szpankowski, Markov Types and Minimax Redundancy for Markov Sources *IEEE Trans. Information Theory*, 50, 1393-1402, 2004.

[17] Eric Kandel, James Schwartz, and Thomas Jessell. *Principles of Neural Science*. Appleton and Lange, 2000.

[18] G. Klir, *Uncertainty and Information: Foundations of Generalized Information Theory*, John Wiley & Sons, New York, 2006.

[19] J. Konorski, A Game-Theoretic Study of CSMA/CA under a Backoff Attack *IEEE/ACM Trans. Networking*, 14, 1167-1178, 2006.

[20] B.O. Kuppers, *Information and the Origin of Life*. The MIT Press, Cambridge, Massachusetts, 1990.

[21] D.G. Luenberger, *Information Science*, Princeton Univ. Press, 2006.

[22] D. Marinescu and M. Marinescu, Quantum Information: a Glimpse at the Strange and Intriguing Future of Information, *The Computer Journal*, 2007.

[23] J. Rissanen, A Universal Data Compression System, *IEEE Trans. Information Theory*, 29, 656–664, 1983.

[24] J. Rissanen, Complexity of Strings in the Class of Markov Sources, *IEEE Trans. Information Theory*, 30, 526–532, 1986.

[25] J. Rissanen, Universal Coding, Information, Prediction, and Estimation, *IEEE Trans. Information Theory*, 30, 629–636, 1984.

[26] J. Rissanen, Fisher Information and Stochastic Complexity, *IEEE Trans. Information Theory*, 42, 40–47, 1996.

[27] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1998.

[28] J. Seidler, *The Science of Information*, WNT, Warszawa, 1982 (in Polish).

[29] A. Shamir, How to Share a Secret, *Communications of the ACM*, 22, 612-613, 1979.

[30] T. Siegfried, *The Bit and the Pendulum: From Quantum Computing to M Theory – The New Physics of Information*, John Wiley & Sons, New York, 2001.

[31] C. Shannon, A Mathematical Theory of Communication, *Bell System Technical Journal*, 27, 379-423 and 623-656, 1948.

[32] C. Shannon. The Lattice Theory of Information. *IEEE Transaction on Information Theory*, 1:105–107, 1953.

[33] W. Szpankowski, On Asymptotics of Certain Recurrences Arising in Universal Coding, *Problems of Information Transmission*, 34, No.2, 142-146, 1998.

[34] W. Szpankowski, A One-to-One Code and Its Anti-redundancy, *2005 International Symposium on Information Theory*, 1526-1528, Adelaide, 2005.

[35] S. Verdù, On Channel Capacity per Unit Cost, *IEEE Trans. Information Theory* 36, 1019-1030, Sep. 1990.

[36] C.F. von Weiszsäcker and E. von Weiszsäcker, Wideraufname der begrifflichen Frage: Was ist Information?, Nova Acta Leopoldina, 206, 1972.

[37] A. D. Wyner, An Upper Bound on the Entropy Series, *Inform. Control*, 20, 176-181, 1972.

[38] P. Young, *The Nature of Information*, Praeger, New York, 1987.