Algorithms, Combinatorics, Information, and Beyond

Plenary Talk at ISIT 2011, St. Petersburg, Russia

April 7, 2012

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

Abstract

Shannon information theory aims at finding fundamental limits for storage and communication, including rates of convergence to these limits. Indeed, many interesting information theoretic phenomena seem to appear in the second order asymptotics. So we first discuss precise analysis of the minimax redundancy that can be viewed as a measure of learnable or useful information. Then we highlight Markov types unveiling some interesting connections to combinatorics of graphical enumeration and linear Diophantine equations. Next we turn our attention to structural compression of graphical objects, proposing a compression algorithm achieving the lower bound represented by the structural entropy. These results are obtained using tools of analytic combinatorics and analysis of algorithms, known also as *analytic information theory*. Finally, we argue that perhaps information theory needs to be broadened if it is to meet today's challenges beyond its original goals (of traditional communication) in biology, economics, modern communication, and knowledge extraction. One of the essential components of this perspective is to continue building foundations in better understanding of temporal, spatial, structural and semantic information in dynamic networks with limited resources. Recently, the National Science Foundation has established the first *Science and Technology Center on Science of Information* (CSoI) to address these challenges and develop tools to move beyond our current understanding of information flow in communication and storage systems.

1 Introduction

It is widely accepted that the information revolution started in 1948 with the publication of Shannon "A Mathematical Theory of Communication". It not only inaugurated a new research field, that of information theory, but also paved the way to today's technological advances in storage and communication such as CDs, iPod, DVD and the internet. Shannon accomplished it all by first introducing a mathematical definition of information that quantifies the extent to which a recipient of data can reduce its statistical uncertainty, and then formulating two fundamental results giving us a lower bound for compression and an upper bound for reliable communication. Furthermore, Shannon declared "these semantic aspects of communication are irrelevant", somewhat abandoning his own dictum in the rate distortion theory (e.g., the distortion measure of audio is incompatible with image compression).

In this article we shall follow another Shannon commandment [66] "it is hardly to be expected that one single concept of information would satisfactorily account for (all) possible

*The work of this author was supported by NSF Science & Technology Center Grant CCF-0939370, NSF Grant CCF-0830140, AFOSR Grant FA8655-11-1-3076, and NSA Grant H98230-11-1-0141.

applications". So we shall argue that information theory may benefit by expanding its original goals to meet today's challenges in biology, economics, modern communication, and knowledge extraction from massive datasets (see also [1]). For this to happen more foundational work in better understanding of temporal, spatial, structural and semantic information is essential.

We are all aware of Shannon warning in his "bandwagon" paper [67] where he thundered "Information theory has, in the last few years, become something of a scientific bandwagon." It is no wonder that some developments in the 50's irked Shannon. Let us just look at the early application of information theory, say to biology. Henry Quastler launched information theory in biology in 1949 (just a year after Shannon's landmark paper and four years before the inception of molecular biology shaped by the work of Crick and Watson) in the paper written together with Dancoff "The Information Content and Error Rate of Living Things". Continuing this effort, Quastler organized two symposiums on "Information Theory in Biology". These attempts were rather unsuccessful as argued by Henry Linschitz [41], who pointed out that there are difficulties in defining information "of a system composed of functionally interdependent units and channel information (entropy) to "produce a functioning cell". To be fair, we need to point out that in 70's Manfred Eigen, Nobel laureate in biochemistry opined, "the differentiable characteristic of the living systems is information. Information assures the controlled reproduction of all constituents, thereby ensuring conservation of viability. Information theory, pioneered by Claude Shannon, cannot answer this question ... in principle, the answer was formulated 130 years ago by Charles Darwin." Eigen's challenge was picked up recently in two new special issues [20, 53] on information theory in molecular biology and neuroscience. The editorial of [20] concludes: "Information Theory is firmly integrated in the fabric of neuroscience research, and a progressively wider range of biological research in general, and will continue to play an important role in these disciplines."

We are now fifty years after the bandwagon paper. In today's world the dynamic flow of information is around us from biology to modern communication to economy. Many scholars argue to broaden information theory beyond its original goals of point-to-point communication and compression of sequences: Sudan and his collaborators [25, 40] suggest that the meaning of information does start to become relevant whenever there is diversity in the communicating parties and when parties themselves evolve over time. For example, when a computer attempts to communicate with a printer both parties must talk the same language in the same format (i.e., "printer driver"). This leads Sudan and his collaborators to consider communication in the setting where encoder and decoder do not agree a priori on the communication protocols, thus encoder and decoder do not understand each other. Bennett in [5] observes that from the earliest days of information theory it has been appreciated that information is not a good *message value.* He continues to propose that the value of information lies in "parts predictable only with difficulties, things that the receiver could figure out without being told". This led him to define the *logical depth.* However, we still do not have a good understanding of the value of information; particularly, in biology and economics. As a matter of fact, in biology, P. Nurse in his 2008 paper [55] claims that biology is on the crossroad and further advances may be required to understand information flow. In Nurse's own words "focusing on information flow will help to understand better how cells and organisms work . . . and temporal order in cell memory and reproduction are not fully understood." Furthermore, in computer science F. Brooks claims [8]: "we have no theory that gives us a metric for the information embodied in structure . . . this is the most fundamental gap in the theoretical underpinning of information and computer science." Finally, Zeilinger goes even further in [9, 85] claiming that reality and information are two sides of the same coin, that is, they are in a deep sense indistinguishable. In communication it is widely accepted that understanding (value and flow of) temporal information is the key to further advances in computer communication [29] and wireless ad-hoc networks [26].

As the matter of fact, in recent decades the information theory community has been pursuing post-Shannon challenges as witnessed in [17, 20, 26, 29, 33, 45, 53, 58, 54, 79], to mention a few. To continue on this path, we propose two approaches that include short(er)-term and long-term research goals:

(i) **Back off from infinity**: Following Ziv's 1997 Shannon Lecture, we propose to extend Shannon findings to finite size data structures (i.e., graphs, sets, social networks), that is, develop *information theory of data structures* beyond first-order asymptotics. We shall argue (see Section 2) that many interesting information-theoretic phenomena appear in the second-order terms. *Analytic information theory* — which applies complex-analytic tools to information theory — is particularly suited for such investigations. We illustrate it in the next section by studying the minimax redundancy problem.

(ii) **Science of Information**: In general, we endeavor to do some foundational work in structural, temporal, spatial and semantic information in dynamic networks with cooperating users (see also recent panel discussion [1]). we also argue that we need a better understanding of complex systems with representation-invariant information. In Section 3 we describe some attempts towards this goal.

In 2010 the National Science Foundation established the first Science and Technology Center for Science of Information (`http://soihub.org`) "*to advance science and technology through a new quantitative understanding of the representation, communication and processing of information in biological, physical, social and engineering systems.*" The center is located at Purdue University and partner institutions include: Bryn Mawr, Berkeley, Howard, MIT, Princeton, Stanford, Texas A&M, UIUC, and UCSD. Some specific Center goals are to: (i) define core theoretical principles governing transfer of information; (ii) develop meters and methods for information; (iii) apply science of information to problems in physical and social sciences, and engineering; and (iv) offer a venue for multi-disciplinary long-term collaborations.

The plan for the paper is as follows. In the next section we discuss the maximal minimax redundancy for memoryless, Markovian, and renewal sources solved by analytic and combinatorial methods. In Section 3 we present a few problems illustrating broader science of information. In particular, we offer some new results on graphical compression as an illustration of structural information.

2 Analytic Information Theory

Jacob Ziv in his 1997 Shannon Lecture presented compelling arguments for "backing off" from first-order asymptotics in order to predict the behavior of real systems with finite length description. To overcome these difficulties, the so called *non-asymptotic analysis*, in which lower and upper bounds are established with controllable error terms, becomes quite popular. However, we argue that developing *full asymptotic expansions* and more precise analysis may be even more desirable. Furthermore, following Hadamard's precept[1], we propose to study information theory problems using techniques of complex analysis[2] such as generating functions, combinatorial calculus, Rice's formula, Mellin transform, Fourier series, sequences distributed modulo 1, saddle point methods, analytic poissonization and depoissonization, and singularity analysis [76].

[1] The shortest path between two truths on the real line passes through the complex plane.

[2] Andrew Odlyzko wrote: "Analytic methods are extremely powerful and when they apply, they often yield estimates of unparalleled precision."

This program, which applies complex-analytic tools to information theory, constitutes analytic information theory.

Analytic information theory can claim some successes in the last decade. We mention a few: proving in the negative the Wyner-Ziv conjecture regarding the longest match [72, 73]; establishing Ziv's conjecture regarding the distribution of the number of phrases in the LZ'78 compression scheme [35, 39]; showing the right order of the LZ'78 redundancy [62, 49]; disproving the Steinberg-Gutman conjecture regarding lossy pattern matching compression schemes [50, 84, 46]; establishing precise redundancy of Huffman's code [75] and redundancy of a fixed-to-variable no prefix free code [77]; deriving precise asymptotics of minimax redundancy for memoryless sources [81, 74, 78], Markov sources [59, 37] and renewal sources [23, 21]; precise analysis of variable-to-fixed codes such as Tunstall and Khodak codes [22]; designing and analyzing error resilient Lemple-Ziv'77 data compression scheme [48], and finally establishing entropy of hidden Markov processes [64] and the noisy constrained capacity [30, 38].

In this section, we illustrate the power of analytic information theory on a few examples taken from the analysis of the minimax redundancy and enumeration of Markov types. First, however, we interpret minimax redundancy as a measure of learnable or useful information capturing regularity properties of an object.

2.1 Learnable/Useful Information and Redundancy

One of the fundamental questions of information theory and statistical inference probes how much "useful or learnable information" one can actually extract from a given data set. To shed some light on this problem, let a binary sequence $x^n = x_1 \dots x_n$ be given. We would like to understand how much useful information, structure, regularity, or summarizing properties are in x^n. For example, for a binary sequence the number of ones is a regularity property, the positions of ones are not. Let in general S be such a summarizing property. We can describe it in two parts. First, we describe the set S, and then the location of x^n in S that requires $\log |S|$ bits (the latter is a good measure of the string complexity). We denote by $I(S)$ the number of bits describing it. Usually, S can be described in many ways, however, one should choose S so that it extracts all relevant information and nothing else. It means we need S that minimizes $I(S)$. We denote such a set as $\hat{S}$ and call it I-sufficient statistic. It makes sense to call $I(\hat{S})$ the *learnable information.*

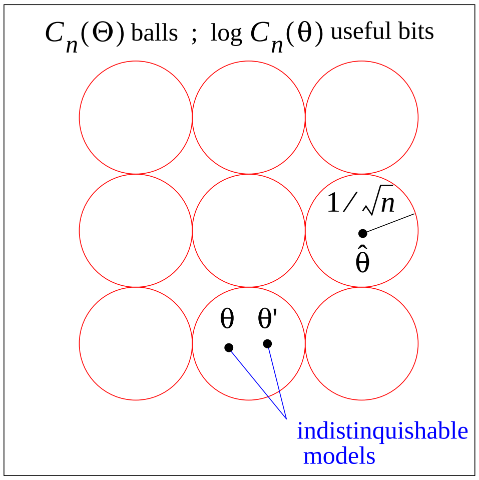


Figure 1: Illustration to $C(\Theta)$

We now consider two concrete measures of learnable information. If $\hat{S}$ is the shortest program on a universal Turing machine, then $I(\hat{S})$ becomes *Kolmogorov-information* [13] $K(\hat{S})$, and $K(x^n) = K(\hat{S}) + \log |\hat{S}|$. For example, if x^n is a binary sequence, we first describe the *type* of x^n (e.g., the number of ones) that requires $O(\log n)$ bits, and then location of x^n within the type which requires $\log \binom{n}{k} \approx nH(k/n)$ bits. While this sounds reasonable, in general Kolmogorov information is not computable, so we need another approach.

We now turn our attention to *computable* useful information contained in a sequence x^n generated by a source belonging to a class of parameterized distributions $\mathcal{M}(\Theta) = \{P_\theta : \ \theta \in \Theta\}$

for some k-dimensional space Θ. We follow here Rissanen [60] and Grunwald [28]. Let $\hat{\theta}(x^n)$ be the maximum likelihood (ML) estimator, that is, $\hat{\theta}(x^n) = \arg\max_{\theta \in \Theta} P_\theta(x^n)$. Observe that for a given sequence x^n, produced either by θ or by θ', we can use $\hat{\theta}(x^n)$ to decide which model generates the data with a small error probability, *provided* these two parameters are far apart in some distance. If these two models, θ and θ' are too close to each others, they are virtually indistinguishable, and they do not introduce any additional useful information. In view of this, it is reasonable to postulate that learnable information about x^n is summarized in the number of *distinguishable distributions* (models), as illustrated in Figure 1. In general, useful information is closely related to *distinguishibility.* In summary, if there are $C_n(\Theta)$ such distinguishable distributions, it is natural to call $I_n(\Theta) = \log C_n(\Theta)$ the *useful information.*

Let us estimate $C_n(\Theta)$ in the MDL (Minimum Description Length) world, as discussed in [3, 28, 60]. As a distance between distributions/models we adopt the Kullback-Leibler (KL) divergence D. Using Taylor expansion around $\hat{\theta}$, we find

$$D(P_{\hat{\theta}}||P_\theta) := \mathbf{E}[\log P_{\hat{\theta}}(X^n)] - \mathbf{E}[\log P_\theta(X^n)] = \frac{1}{2}(\theta - \hat{\theta})^T I(\hat{\theta})(\theta - \hat{\theta}) + o(||\theta - \hat{\theta}||^2), \qquad (1)$$

where $I(\theta) = \{I_{ij}(\theta)\}_{ij}$ is the *Fisher information matrix* defined as

$$I_{ij}(\theta) = -\mathbf{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log P_\theta(X)\right].$$

As a distance we use

$$d_I(\theta, \theta_0) = \sqrt{(\theta - \hat{\theta})^T I(\hat{\theta})(\theta - \hat{\theta})}$$

which is the so called *Mahalanobis distance* [28]. This is a rescaled version of Euclidean distance, and by (1) we have $d_I(\theta, \theta_0) = O(\sqrt{D(\theta||\theta_0)})$. One property of the d_I distance is that the volume V of a ball (ellipsoid) at center θ and radius ε is

$$V(B_I(\theta, \varepsilon)) = 1/\sqrt{\det I(\theta)} V(B(\varepsilon),$$

where $B(\varepsilon)$ is the regular Euclidean ball and $\det I(\theta)$ is the determinant of $I(\theta)$ [3, 28].

To proceed, we need to specify the error probability and distinguishibility. Let $B_{KL}(\theta_0, \varepsilon) = \{\theta : D(\theta||\theta_0) \leq \varepsilon\}$ be the KL-ball or radius ε around θ_0. Observe that the KL-ball $B_{KL}(\theta, \varepsilon)$ becomes $B_I(\theta, \sqrt{\varepsilon})$ ball in the d_I distance. The distinguishability of models depends on the error probability that can be estimated as follows [3] for some $\theta \in \Theta_0$ with $\dim(\Theta_0) = k$ [28]

$$P_\theta(\hat{\theta} \neq \theta) \quad = \quad P_\theta(\arg\min_{\theta \in \Theta_0} D(\hat{\theta}(X^n)||\theta) \neq \theta) \approx P_\theta(\theta(X) \notin B_{KL}(\theta, \varepsilon/n)) \sim 1 - O(\varepsilon^{k/2})$$

for some small $\varepsilon > 0$, where we use the fact that for Markov sources (more generally, for an exponential family of distributions)

$$\log \frac{P_{\hat{\theta}}(x^n)}{P_\theta(x^n)} = n\mathbf{E}_{\hat{\theta}}\left[\log \frac{P_{\hat{\theta}}(X)}{P_\theta(X)}\right] = nD(\hat{\theta}||\theta).$$

We conclude that the number of distinguishable distributions $C_n(\Theta)$ is approximately equal to the *volume* $V_I(\Theta)$ of Θ under distance d_I divided by the volume of the ball size $B_I(\theta, \sqrt{\varepsilon/n})$. In [3] it is proved that

$$V_I(\Theta) = \int_\Theta \sqrt{\det I(\theta)}\, d\theta, \qquad V(B_I(\theta, \sqrt{\varepsilon}) \approx O(\varepsilon^{k/2}/\sqrt{\det I(\theta)}).$$

Setting up the error probability at level $O(1/\sqrt{n})$ as indicated above, we conclude that the number of distinguishable distributions $C_n(\Theta)$ (i.e., the number of centers of the balls $B_I(\theta, \sqrt{\varepsilon})$) is (see [21, 28, 60])

$$C_n(\Theta) = \left(\frac{n}{2\pi}\right)^{k/2} \int_\Theta \sqrt{\det I(\theta)}d\theta + O(1) = \sum_{x^n} \sup_{\theta\in\Theta} P_\theta(x^n) = \inf_{\theta\in\Theta} \max_{x^n} \log \frac{P_{\hat{\theta}}}{P_\theta} \tag{2}$$

where the second equality follows from [28, 59]. In order to justify the last equality we need to turn our attention to the *maximal minimax redundancy.*

Let us begin with a precise information-theoretic definition of the minimax redundancy and its Shtarkov's bounds. Throughout this section, we write $L(C_n, x^n)$ for the length of a fixed-to-variable code $C_n; \mathcal{A}^n \to \{0,1\}^*$ assigned to the source sequence x^n over the alphabet $\mathcal{A} = \{1, 2, \ldots, m\}$ of size m that can be finite or not. In practice, one can only hope to have some knowledge about a *family* of sources $\mathcal{S}$ that generates the data, such as the family of memoryless sources $\mathcal{M}_0$ or Markov sources $\mathcal{M}_r$ of order $r > 0$. Following Davisson [18] and Shtarkov [69], we define the minimax worst-case (maximal) redundancy $R_n^*(\mathcal{S})$ for a family $\mathcal{S}$ as

$$R_n^*(\mathcal{S}) = \min_{C_n} \sup_{P\in\mathcal{S}} \max_{x_1^n} \left[L(C_n, x_1^n) + \log P(x_1^n)\right], \tag{3}$$

where C_n represents a set of prefix codes, and the source $P \in \mathcal{S}$ generates the sequence $x^n = x_1 \ldots x_n$. If we ignore the integer nature of the code length $L(C_n, x^n)$, then we can approximate it by $\log 1/P_\theta$ for some θ. Furthermore, $\log \sup_{P\in\mathcal{S}} P(x^n) = \log(1/P_{\hat{\theta}})$, where $\hat{\theta}$ is the ML estimator, so that

$$R_n^*(\mathcal{S}) = \inf_\theta \max_{x^n} \log \frac{P_{\hat{\theta}}}{P_\theta} + O(1) \tag{4}$$

which is the right-hand side of (2), and therefore $C(\Theta) = R_n^*(\mathcal{S}) + O(1)$.

We still need to justify the last equality in (2). We derive now Shtarkov's bound [69]. Define first the maximum likelihood distribution

$$Q^*(x^n) := \frac{\sup_{P\in\mathcal{S}} P(x^n)}{\sum_{y^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y^n)}.$$

Then observe [21]

$$\begin{aligned}
R_n^*(\mathcal{S}) &= \min_{C_n} \sup_{P\in\mathcal{S}} \max_{x^n} (L(C_n, x^n) + \log P(x^n)) \\
&= \min_{C_n} \max_{x^n} \left(L(C_n, x^n) + \sup_{P\in\mathcal{S}} \log P(x^n)\right) \\
&= \min_{C_n} \max_{x^n} \left(L(C_n, x^n) + \log Q^*(x^n)\right) + \log \sum_{y^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y^n) \\
&= R_n^{GS}(Q^*) + \log \sum_{y^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y^n) = \log \sum_{y^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y^n) + O(1)
\end{aligned}$$

where $0 < R_n^{GS}(Q^*) \leq 1$ is the redundancy of the optimal generalized Shannon code (see [21]). Therefore, ignoring again the integer constraint (i.e., setting $R_n^{GS}(Q^*) = 0$) and using (4) rather than (3) we arrive at

$$\sum_{x^n} \sup_{\theta\in\Theta} P_\theta(x^n) = \inf_{\theta\in\Theta} \max_{x^n} \log \frac{P_{\hat{\theta}}}{P_\theta} = R_n^*(\mathcal{S})$$

which establishes the right-hand side of (2). From now on, we assume that $R^*(\mathcal{S}) = \log D_{n,m}(\mathcal{S})$ where

$$D_{n,m}(\mathcal{S}) = \sum_{x^n \in \mathcal{A}^n} \sup_{P \in \mathcal{S}} P(x^n). \tag{5}$$

The $O(1)$ term in (4) can be computed for finitely parameterized sources as in [21], but we will not elaborate on it here.

In summary, useful or learnable information is closely related to the minimax redundancy $R_n^*(\mathcal{S})$ which can be viewed as a measure of certain regularity properties of a source (regularity beyond the randomness/complexity expressed by the entropy). Next, using analytic tools we estimate asymptotically the minimax redundancy for various classes of sources such as memoryless for finite and infinite alphabets, renewal sources, and Markov sources. When discussing Markov sources, we rather turn our attention to combinatorial aspects of Markov types.

2.2 Minimax Redundancy for Memoryless Sources

In this section we study the minimax redundancy for a class of memoryless sources over finite and infinite alphabet of size m. We follow here [74]. Observe that $D_{n,m} := D_{n,m}(\mathcal{M}_0)$ defined in (5) takes the form

$$D_{n,m} = \sum_{k_1+\cdots+k_m=n} \binom{n}{k_1,\ldots,k_m} \left(\frac{k_1}{n}\right)^{k_1} \cdots \left(\frac{k_m}{n}\right)^{k_m}, \tag{6}$$

where k_i is the number of times symbol $i \in \mathcal{A}$ occurs in a string of length n. Indeed, observing that $P(x^n) = p_1^{k_1} \cdots p_m^{k_m}$ where p_i are *unknown* parameters θ representing the probability for symbol $i \in \mathcal{A}$, we proceed as follows

$$\begin{aligned} D_n(\mathcal{M}_0) &= \sum_{x_1^n} \sup_{P(x_1^n)} P(x_1^n) \\ &= \sum_{x_1^n} \sup_{p_1,\ldots,p_m} p_1^{k_1} \cdots p_m^{k_m} \\ &= \sum_{k_1+\cdots+k_m=n} \binom{n}{k_1,\ldots,k_m} \sup_{p_1,\ldots,p_m} p_1^{k_1} \cdots p_m^{k_m} \\ &= \sum_{k_1+\cdots+k_m=n} \binom{n}{k_1,\ldots,k_m} \left(\frac{k_1}{n}\right)^{k_1} \cdots \left(\frac{k_m}{n}\right)^{k_m}, \end{aligned}$$

where the last line follows from

$$\sup_{P(x_1^n)} P(x_1^n) = \sup_{p_1,\ldots,p_m} p_1^{k_1} \cdots p_m^{k_m} = \left(\frac{k_1}{n}\right)^{k_1} \cdots \left(\frac{k_m}{n}\right)^{k_m}.$$

We should point out that (6) has a form that re-appears in the redundancy analysis of other sources. Indeed, the summation is over tuples $\mathbf{k} = (k_1, \ldots, k_m)$ representing a (memoryless) *type* (cf. Section 2.4) and under the sum the first term $\binom{n}{k_1,\ldots,k_m}$ counts the number of sequences x^n of the same type while the second term is the maximum likelihood distribution.

It is argued in [74] that the asymptotics of such a sum can be analyzed through its so-called *tree-like generating function* defined as

$$D_m(z) = \sum_{n=0}^{\infty} \frac{n^n}{n!} D_{n,m} z^n.$$

Here, we will follow the same methodology and employ the convolution formula for tree-like generating functions (cf. [76]). Observe that $D_m(z)$ relates to another tree-like generating function defined as

$$B(z) = \sum_{k=0}^{\infty} \frac{k^k}{k!} z^k.$$

This function, in turn, can be shown to be (cf. [76]) $B(z) = (1 - T(z))^{-1}$ for $|z| < e^{-1}$, where $T(z) = \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} z^k$ is the well-known *tree function* — that counts the number of rooted labeled trees on n vertices [24] — satisfying the implicit equation

$$T(z) = ze^{T(z)} \tag{7}$$

with $|T(z)| < 1$. The convolution formula [76] applied to (6) yields

$$D_m(z) = [B(z)]^m - 1. \tag{8}$$

Consequently, $D_{n,m} = \frac{n!}{n^n} [z^n] \, [B(z)]^m$ where $[z^n] f(z)$ denotes the coefficient of z^n in $f(z)$.

Defining $\beta(z) = B(z/e)$, $|z| < 1$, noticing that $[z^n]\beta(z) = e^{-n}[z^n]B(z)$, and applying Stirling's formula, (8) yields

$$D_{n,m} = \sqrt{2\pi n} \left(1 + O(n^{-1})\right) [z^n] \, [\beta(z)]^m \, . \tag{9}$$

Thus, it suffices to extract asymptotics of the coefficient at z^n of $[\beta(z)]^m$, for which a standard tool is Cauchy's coefficient formula [24, 76], that is,

$$[z^n][\beta(z)]^m = \frac{1}{2\pi i} \oint \frac{\beta^m(z)}{z^{n+1}} dz \tag{10}$$

where the integration is around a closed path containing $z = 0$ inside which $\beta^m(z)$ is analytic. However, asymptotic evaluation of the above depends whether m is finite or is a function of n. We consider these two cases next.

2.2.1 Finite Alphabet Size

First we assume that the size of the alphabet m is finite and does not depend on n. This case was analyzed in [74] (see also [81, 82]). To evaluate the integral in (10) we apply Flajolet and Odlyzko *singularity analysis* [24, 76] because $[\beta(z)]^m$ has only algebraic singularities. Indeed, using (7) it can be shown that the singular expansion of $\beta(z)$ around its singularity $z = 1$ is [12]

$$\beta(z) = \frac{1}{\sqrt{2(1-z)}} + \frac{1}{3} - \frac{\sqrt{2}}{24}\sqrt{(1-z)} + O(1-z).$$

From [24, 76] we know that

$$[z^n](1-z)^{-\alpha} \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)}, \quad \alpha \notin \{0, -1, -2, \ldots\}.$$

This is illustrated in Figure 2. The singularity analysis then yields the minimax redundancy [74]

$$R^*_{n,m} \;:=\; \log D_{n,m} = \frac{m-1}{2} \log\left(\frac{n}{2}\right) + \log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) + \frac{\Gamma(\frac{m}{2}) m \log e}{3\Gamma(\frac{m}{2} - \frac{1}{2})} \cdot \frac{\sqrt{2}}{\sqrt{n}} + O\left(\frac{1}{n}\right) \tag{11}$$

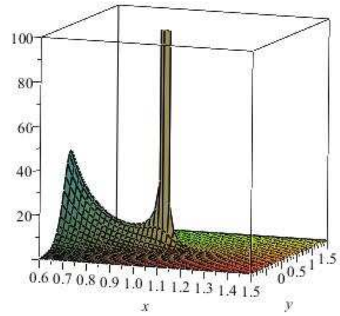


Figure 2: Singularity analysis

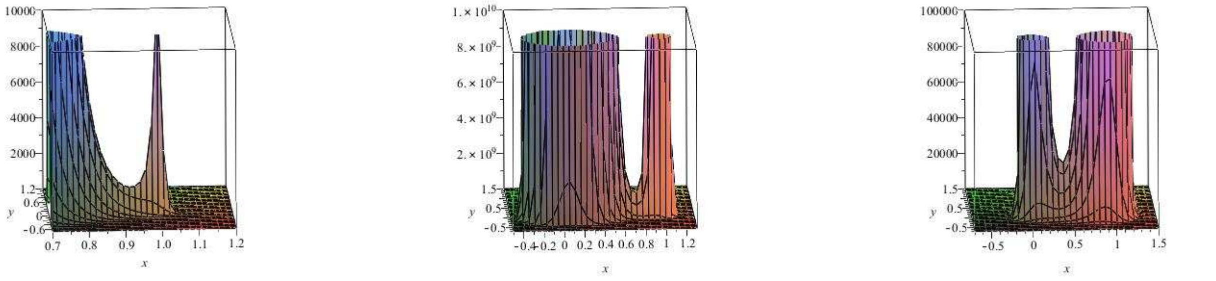
for large n and fixed m, where Γ is the Euler gamma function. We conclude that the first term above coincides with Rissanen's lower bound: we pay a penalty of $\log n/2$ per unknown parameter.

2.2.2 Unbounded Alphabet

Now we assume that the alphabet size is unknown and *unbounded.* In fact, it may depend on n. When m grows with n, the singularity analysis *does not* apply because $\beta^m(z)$ grows exponentially with n. The growth of $\beta^m(z)$ determines that the *saddle point method* [24, 76], which we briefly review next, can be applied to (10). We will restrict our attention to a special case of the method, where the goal is to obtain an asymptotic approximation of

$$D_{n,m} = \sqrt{2\pi n}\frac{1}{2\pi i}\oint \frac{\beta(z)^m}{z^{n+1}}dz = \sqrt{2\pi n}\frac{1}{2\pi i}\oint e^{g(z)}dz,$$

where $g(z) = m\ln\beta(z) - (n+1)\ln z$. For example, when $m = n+1$ the function under the integral grows as $\exp\left((n+1)[\ln\beta(z) - \ln z]\right)$ which becomes very large around z where $\ln\beta(z) - \ln z$ is maximized, and almost negligible everywhere else. This determines the asymptotics.



$m = o(n)$ $m = n$ $n = o(m)$

Figure 3: Illustration of the saddle point method for Theorem 1.

In general for any m and n, the saddle point z_0 is a solution of $g'(z_0) = 0$, which yields

$$g(z) = g(z_0) + \frac{1}{2}(z - z_0)^2 g''(z_0) + O(g'''(z_0)(z - z_0)^3).$$

Under mild conditions (see Table 8.4 in [76]), satisfied by our $g(z)$ (e.g., z_0 is real and unique), the saddle point method leads to

$$D_{n,m} = \sqrt{2\pi n}\frac{e^{g(z_0)}}{\sqrt{2\pi|g''(z_0)|}} \times \left(1 + O\left(\frac{g'''(z_0)}{(g''(z_0))^\rho}\right)\right),$$

for some $\rho < 3/2$. In our case, the saddle point z_0 varies from near 1 to near 0 depending on the relation between n and m as illustrated in Figure 3. It turns out that three cases must be considered: $m = o(n)$ (the saddle point $z_0 \approx 1$), $m = O(n)$ (saddle point $0 < z_0 < 1$), and the case $n = o(m)$ (in this case $z_0 \approx 0$).

The following result is proved in [78] (see also [56])

Theorem 1 (Szpankowski and Weinberger, 2010)**.** *For memoryless sources $\mathcal{M}_0$ over an m-ary alphabet, where $m \to \infty$ as n grows, the minimax worst-case redundancy behaves asymptotically as follows:*
(i) *For $m = o(n)$*

$$R^*_{n,m} \;=\; \frac{m-1}{2}\log\frac{n}{m} + \frac{m}{2}\log e + \frac{m\log e}{3}\sqrt{\frac{m}{n}} - \frac{1}{2} - \frac{\log e}{4}\sqrt{\frac{m}{n}} + O\left(\frac{m^2}{n} + \frac{1}{\sqrt{m}}\right). \quad (12)$$

(ii) *For $m = \alpha n + \ell(n)$, where α is a positive constant and $\ell(n) = o(n)$,*

$$R^*_{n,m} \;=\; n\log B_\alpha + \ell(n)\log C_\alpha - \log\sqrt{A_\alpha} - \frac{\ell(n)^2\log e}{2n\alpha^2 A_\alpha} + O\left(\frac{\ell(n)^3}{n^2} + \frac{\ell(n)}{n} + \frac{1}{\sqrt{n}}\right), \quad (13)$$

where

$$C_\alpha := \frac{1}{2} + \frac{1}{2}\sqrt{1+\frac{4}{\alpha}}, \qquad A_\alpha := C_\alpha + \frac{2}{\alpha}, \qquad B_\alpha = \alpha C_\alpha^{\alpha+2} e^{-\frac{1}{C_\alpha}}.$$

(iii) *For $n = o(m)$*

$$R^*_{n,m} = n\log\frac{m}{n} + \frac{3}{2}\frac{n^2}{m}\log e - \frac{3}{2}\frac{n}{m}\log e + O\left(\frac{1}{\sqrt{n}} + \frac{n^3}{m^2}\right). \quad (14)$$

In summary, we conclude that for finite m and $m = o(n)$ the minimax redundancy, representing useful information embodied in regularity properties of a sequence, grows like $(m-1)/2 \times \log(n/m)$. This coincides with Rissanen's lower bound. However, for $m = O(n)$ the minimax redundancy grows linearly with n while for m growing faster than n the growth of the minimax redundancy is $n\log(m/n)$.

2.3 Minimax Redundancy for Renewal Sources

Let us continue our analytic *extravaganza* and consider non-finitely parameterized sources before we return to Markovian sources over finite alphabets in the next section. We study here the so called *renewal sources* first introduced in 1996 by Csiszár and Shields [15]. Such a source is defined as follows:

- Let $T_1, T_2 \ldots$ be a sequence of i.i.d. positive-valued random variables with distribution $Q(j) = \Pr\{T_i = j\}$.
- The process $T_0, T_0 + T_1, T_0 + T_1 + T_2, \ldots$ is a renewal process.
- In a binary renewal sequence the positions of the 1's are at the renewal epochs $T_0, T_0+T_1, \ldots$ with runs of zeros of lengths $T_1 - 1$, $T_2 - 1$, ... in between the 1's.
- The process starts with $x_0 = 1$.

We follow here the analysis presented in [23]. A sequence generated by such a source becomes

$$x_0^n = 10^{\alpha_1}10^{\alpha_2}1\cdots 10^{\alpha_n}1\underbrace{0\cdots 0}_{k^*}$$

where k_m is the number of i such that $\alpha_i = m$. Then

$$P(x_1^n) = [Q(0)]^{k_0}[Q(1)]^{k_1}\cdots[Q(n-1)]^{k_{n-1}}\Pr\{T_1 > k^*\}.$$

The last term introduces some difficulties in finding the maximum likelihood distribution, but it can be proved that the minimax redundancy $R_n^*(\mathcal{R}_0) = \log D_n(\mathcal{R}_0)$ of the renewal source $\mathcal{R}_0$ satisfies

$$r_{n+1} - 1 \le D_n(\mathcal{R}_0) \le \sum_{m=0}^{n} r_m$$

where $r_n = \sum_{k=0}^{n} r_{n,k}$ and

$$r_{n,k} \quad = \quad \sum_{\mathcal{I}(n,k)} \binom{k}{k_0 \cdots k_{n-1}} \left(\frac{k_0}{k}\right)^{k_0} \left(\frac{k_1}{k}\right)^{k_1} \cdots \left(\frac{k_{n-1}}{k}\right)^{k_{n-1}}. \tag{15}$$

Above $\mathcal{I}(n,k)$ is is the integer partition of n into k terms, i.e.,

$$n = 1k_0 + 2k_1 + \cdots + nk_{n-1}, \quad k = k_0 + \cdots + k_{n-1}.$$

Since r_n is too difficult to analyze, we rather study $s_n = \sum_{k=0}^{n} s_{n,k}$ where

$$s_{n,k} = e^{-k} \sum_{\mathcal{P}(n,k)} \frac{k^{k_0}}{k_0!} \cdots \frac{k^{k_{n-1}}}{k_{n-1}!}, \qquad \frac{r_{n,k}}{s_{n,k}} = \frac{k!}{k^k e^{-k}}$$

since

$$S(z,u) = \sum_{k,n} s_{n,k} (u/e)^k z^n = \sum_{\mathcal{P}_{n,k}} z^{1k_0 + 2k_1 + \cdots + nk_{n-1}} \left(\frac{u}{e}\right)^{k_0 + \cdots + k_{n-1}} \frac{k^{k_0}}{k_0!} \cdots \frac{k^{k_{n-1}}}{k_{n-1}!} = \prod_{i=1}^{\infty} \beta(z^i u)$$

where $\beta(z) = B(z/e)$ is defined in the previous section.

To compare s_n to r_n, we introduce the random variable K_n as follows

$$\Pr\{K_n = k\} = \frac{s_{n,k}}{s_n}.$$

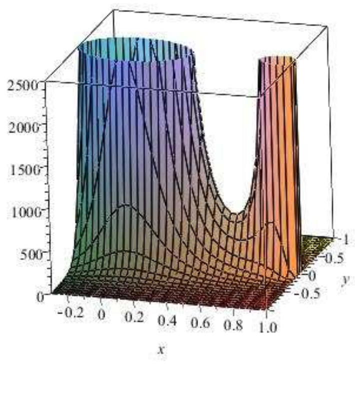


Figure 4: Saddle Point

Stirling's formula yields

$$\frac{r_n}{s_n} = \sum_{k=0}^{n} \frac{r_{n,k}}{s_{n,k}} \frac{s_{n,k}}{s_n} \quad = \quad \mathbf{E}[(K_n)! K_n^{-K_n} e^{K_n}]$$

$$= \quad \mathbf{E}[\sqrt{2\pi K_n}] + O(\mathbf{E}[K_n^{-\frac{1}{2}}]).$$

Thus

$$r_n = s_n \mathbf{E}[\sqrt{2\pi K_n}](1 + o(1)) = s_n \sqrt{2\pi \mathbf{E}[K_n]}(1 + o(1)).$$

To understand probabilistic behavior of K_n, we apply sophisticated tools of analytic combinatorics such as Mellin transform and the saddle point [24, 76]. In particular, we must evaluate $[z^n]S(z,1)$ by the saddle point that leads to the following

$$s_n = [z^n]S(z,1) = [z^n] \exp\left(\frac{c}{1-z} + a \log \frac{1}{1-z}\right)$$

which is illustrated in Figure 4. We prove in [23] the following.

Lemma 1. *Let $\mu_n = \mathbf{E}[K_n]$ and $\sigma_n^2 = \mathbf{Var}(K_n)$. Then*

$$\mu_n = \frac{1}{4}\sqrt{\frac{n}{c}}\log\frac{n}{c} + o(\sqrt{n}), \qquad \sigma_n^2 = O(n\log n) = o(\mu_n^2),$$

where $c = \pi^2/6 - 1$, $d = -\log 2 - \frac{3}{8}\log c - \frac{3}{4}\log \pi$.

This leads to our final result proved in [23].

Theorem 2 (Flajolet and Szpankowski, 1998)**.** *We have the following asymptotics*

$$\begin{aligned} s_n &\sim \exp\left(2\sqrt{cn} - \frac{7}{8}\log n + O(1)\right), \\ \log r_n &= \frac{2}{\log 2}\sqrt{cn} - \frac{5}{8}\log n + \frac{1}{2}\log\log n + O(1). \end{aligned}$$

that yields

$$R_n^*(\mathcal{R}_0) = \frac{2}{\log 2}\sqrt{cn} + O(\log n).$$

where $c = \frac{\pi^2}{6} - 1 \approx 0.645$.

In passing we should point out that the renewal source technically is reminiscent of the memoryless sources with unbounded alphabet (cf (3) and (15)). The analysis of renewal sources is, however, much more sophisticated.

2.4 Markov Minimax Redundancy and Markov Types

In this section, we return to the finite size alphabet $\mathcal{A} = \{1, \ldots, m\}$ but now we consider a class $\mathcal{M}_1$ of Markovian sources of order $r = 1$. More precisely, the probability of a sequence x^n is given by

$$P(x^n) = P(x_1)\prod_{i,j=1}^{m} p_{ij}^{k_{ij}}$$

where k_{ij} is the number of pairs $(i,j) \in \mathcal{A}^2$ in x^n, p_{ij} are the (unknown) transition probabilities while $P(x_1)$ is the initial probability. Then the minimax redundancy (ignoring again the integer nature of coding) is [37]

$$D_n(\mathcal{M}_1) = \sum_{x_1^n}\sup_P\ P(x^n) = \sum_{\mathbf{k}\in\mathcal{Q}_n(m)} |\mathcal{T}_n(\mathbf{k})| \left(\frac{k_{11}}{k_1}\right)^{k_{11}} \cdots \left(\frac{k_{mm}}{k_m}\right)^{k_{mm}}, \tag{16}$$

where $\mathcal{Q}_n(m)$ denotes a set of *Markov types* discussed in the sequel, and $\mathcal{T}_n(\mathbf{k}) := \mathcal{T}_n(x_1^n)$ is the number of sequences of the same Markov type represented by the frequency count matrix $\mathbf{k} = \{k_{ij}\}_{i,j\in\mathcal{A}^2}$. The *frequency matrix* $\mathbf{k}$, which we also write $[k_{ij}]$, satisfies two important properties

$$\sum_{i,j\in\mathcal{A}} k_{ij} = n - 1, \tag{17}$$

and additionally for any $i \in \mathcal{A}$ [37, 83]

$$\sum_{j=1}^{m} k_{ij} = \sum_{j=1}^{m} k_{ji} + \delta(x_1 = i) - \delta(x_n = i), \quad \forall i \in \mathcal{A}, \tag{18}$$

$$\mathbf{k} = \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$$

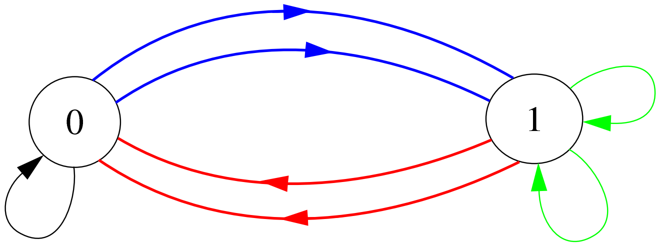


Figure 5: A frequency matrix and its corresponding Eulerian graph.

where $\delta(A) = 1$ when A is true and zero otherwise. The last property is called the *flow conservation property* and is a consequence of the fact that the number of pairs starting with symbols $i \in \mathcal{A}$ must be equal to the number of pairs ending with symbol $i \in \mathcal{A}$ with the possible exception of the first and last pairs. To avoid this exception, hereafter we focus on *cyclic* strings in which the first element x_1 follows the last x_n. For such cyclic strings the frequency matrix $\mathbf{k}$ satisfies a simplified system of linear equations, namely

$$\sum_{i,j\in\mathcal{A}} k_{ij} = n, \tag{19}$$

$$\sum_{j=1}^{m} k_{ij} = \sum_{j=1}^{m} k_{ji}, \qquad \forall\, i \in \mathcal{A}. \tag{20}$$

Such integer matrices $\mathbf{k}$ will be called *balanced frequency matrices* or simply balanced matrices. We also call (20) the "conservation law" equation or simply the *balanced boundary condition* (BBC). We denote by $\mathcal{F}_n(m)$ the set of nonnegative integer solutions of (19) and (20).

We are now ready to define cyclic Markov types. Two cyclic sequences have the same (cyclic) Markov type if they have the same empirical distribution

$$P(x^n) = \prod_{i,j\in\mathcal{A}} p_{ij}^{k_{ij}}.$$

Thus, we assume the initial condition is a cyclic one. We denote by $\mathcal{P}_n(m)$ the set of cyclic Markov types and enumerate them by comparing them to the cardinality of $\mathcal{F}_n(m)$, and also to the set of Markov types $\mathcal{Q}_n(m)$ over linear strings. In passing, we should point out that for a given sequence x^n, the *type class* is defined as

$$\mathcal{T}_n(x^n) = \{y^n : \; P(x^n) = P(y^n)\}$$

for all empirical distributions P_{x^n} in a given model class. Clearly, $\bigcup_{x^n} \mathcal{T}_n(x^n) = \mathcal{A}^n$, and $|\mathcal{T}_n(x^n)|$ counts the number of sequences of the same type as x^n; it is required to estimate the minimax redundancy for Markov sources as shown in (16).

Our goal is to enumerate the number of cyclic Markov types $|\mathcal{P}_n(m)|$ that from now on we simply call Markov types. Enter combinatorics: we shall show that the number of Markov types is asymptotically equivalent to estimating; (i) the number of the balanced frequency matrices, (ii) the number of integer solutions $|\mathcal{F}_n(m)|$ of a system of *linear Diophantine equations* (19)–(20), and finally (iii) the number of connected Eulerian multigraphs, as defined next. To see the latter, we present another characterization of Markov types. Let us define a directed multigraph $G = (V, E)$ with the set of vertices $V = \mathcal{A}$ and k_{ij} edges between vertices $i, j \in \mathcal{A}$. For

$\mathcal{A} = \{0, 1\}$ such a graph is shown in Figure 5. Then, as already observed in [6, 27, 37], the number of sequences of a given type $\mathbf{k}$, i.e., $|\mathcal{T}(\mathbf{k})|$, is equal to the number of Eulerian cycles in G. On the other hand, the number of types $|\mathcal{P}_n(m)|$ coincides with the number of Eulerian digraphs $G = (V, E)$ such that $V \subseteq \mathcal{A}$ and $|E| = n$ (here $V \subseteq \mathcal{A}$ since there may be sequences composed of only *some* symbols of the alphabet). The point we emphasize is that G may be defined over a *subset* of $\mathcal{A}$, as shown in the next Figure 6 (i.e., there may be some isolated vertices).

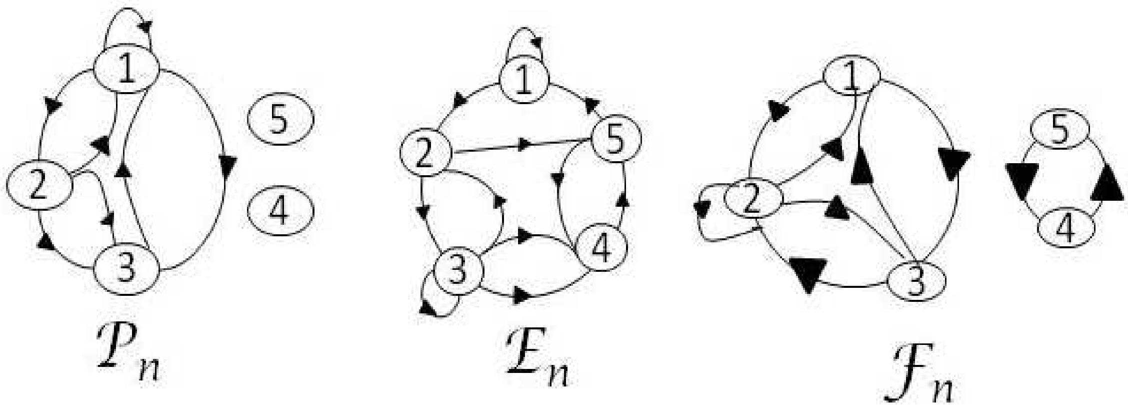


Figure 6: Examples of graphs belonging to $\mathcal{P}_7(5)$, $\mathcal{E}_{11}(5)$ and $\mathcal{F}_9(5)$ sets.

Let us explore further these two sets $\mathcal{P}_n(m)$ and $\mathcal{F}_n(m)$ in the language of graphs. In fact, we need to introduce another set. We denote it by $\mathcal{E}_n(m)$, the set of connected Eulerian digraphs on $\mathcal{A}$; the middle of Figure 6 shows an example of a graph in this set. Finally, the set $\mathcal{F}_n(m)$ can be viewed as the set of digraphs G with $V(G) = \mathcal{A}$, $|E(G)| = n$ and satisfying the flow conversation property (in-degree equals out-degree). We call such graphs *conservative digraphs.* Observe that a graph in $\mathcal{F}_n(m)$ may consist of several connected (not communicating) Eulerian digraphs, as shown in the third example in Figure 6.

There is a simple relation between $|\mathcal{E}_n(m)|$ and $|\mathcal{P}_n(m)|$. Indeed,

$$|\mathcal{P}_n(m)| = \sum_k \binom{m}{k} |\mathcal{E}_n(k)| \tag{21}$$

since there are $\binom{m}{k}$ ways to choose $m - k$ isolated vertices in $\mathcal{P}_n(m)$. Now, observe that a conservative digraph may have several connected components. Each connected component is either a connected Eulerian digraph or an isolated node without an edge. This leads to

$$|\mathcal{F}_n(m)| = |\mathcal{E}_n(m)| + \sum_{i=2}^{m} \sum_{\mathcal{A}=\mathcal{A}_1\cup\cdots\cup\mathcal{A}_i} \sum_{n_1+\cdots+n_i=n} \prod_{j=1}^{i} |\mathcal{E}_{n_j}(\mathcal{A}_j)| \tag{22}$$

where the sum is over all (unordered) set partitions $\mathcal{A} = \mathcal{A}_1 \cup \cdots \cup \mathcal{A}_i$ into $i \geq 2$ (nonempty) parts with n_j edges in each di-subgraph $\mathcal{E}_{n_j}(\mathcal{A}_j)$ over $\mathcal{A}_j$ vertices. Observe that every set partition $\mathcal{A} = \mathcal{A}_1 \cup \cdots \cup \mathcal{A}_i$ with $|\mathcal{A}_j| = m_j > 0$ is a partition of $\mathcal{A}$ into i *distinguished* subsets of cardinality m_j. In fact, using the so called *exponential formula* [24] (page 118) we may conclude even more, namely [34]

$$|\mathcal{F}_n(m)| = |\mathcal{E}_n(m)| + \sum_{i=2}^{m} \frac{1}{i!} \sum_{m_1+\cdots+m_i=m} \binom{m}{m_1 \cdots m_i} \sum_{n_1+\cdots n_i=n} \prod_{j=1}^{i} |\mathcal{E}_{n_j}(m_j)|.$$

A direct consequence of this is the following asymptotic equivalence [34].

Lemma 2. *The following holds for all $m \geq 2$ and $n \to \infty$*

$$|\mathcal{F}_n(m)| = |\mathcal{P}_n(m)| + O(2^m m^3 n^{m^2-3m+3}). \tag{23}$$

In view of the above we need to enumerate the number of solutions $|\mathcal{F}_n(m)|$ of the system of linear Diophantine equations (19)–(20). Again, we accomplish it by analytic methods. Let

$$F_m^*(z) = \sum_{n\geq 0} |\mathcal{F}_n(m)| z^n.$$

However, to find $F_m^*(z)$ we need to evaluate a more complicated generating function that enumerates all balanced matrices, that is,

$$F_m^*(\mathbf{z}) = \sum_{\mathbf{k}\in\mathcal{F}_n(m)} \mathbf{z}^{\mathbf{k}},$$

where $\mathbf{z}^{\mathbf{k}} := \prod_{ij} z_{ij}^{k_{ij}}$. Notice that the summation is over all balances matrices $\mathbf{k} \in \mathcal{F}_n(m)$. This is a daunting task, but we can easily compute the above generating function if the summation is over *all* matrices (satisfying only (19)). Indeed,

$$F_m(\mathbf{z}) = \sum_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} = \prod_{ij} (1 - z_{ij})^{-1}. \tag{24}$$

The remaining problem is to translate $F_m(\mathbf{z})$ into $F_m^*(\mathbf{z})$. This is presented in the next lemma, where we consider a multivariate generating functions $G(\mathbf{z}) = \sum_{\mathbf{k}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}}$ and $G^*(\mathbf{z}) = \sum_{\mathbf{k}\in\mathcal{F}} g_{\mathbf{k}}\mathbf{z}^{\mathbf{k}} = \sum_{n\geq 0}\sum_{\mathbf{k}\in\mathcal{F}_n(m)} g_{\mathbf{k}}\mathbf{z}^{\mathbf{k}}$ over general sequences $g_{\mathbf{k}}$ indexed by matrices $\mathbf{k}$. The following was proved in [37].

Lemma 3. *Let $G(\mathbf{z}) = \sum_{\mathbf{k}} g_{\mathbf{k}}\mathbf{z}^{\mathbf{k}}$ be the generating function of a complex matrix $\mathbf{z}$. Then*

$$G^*(\mathbf{z}) := \sum_{n\geq 0}\sum_{\mathbf{k}\in\mathcal{F}_n} g_{\mathbf{k}}\mathbf{z}^{\mathbf{k}} \quad = \quad \left(\frac{1}{2\mathbf{i}\pi}\right)^m \oint \frac{dx_1}{x_1} \cdots \oint \frac{dx_m}{x_m} G\left(\left[z_{ij}\frac{x_j}{x_i}\right]\right)$$

with the convention that the ij-th coefficient of the matrix $[z_{ij}\frac{x_j}{x_i}]$ is $z_{ij}\frac{x_j}{x_i}$, and $\mathbf{i} = \sqrt{-1}$. In other words, $[z_{ij}\frac{x_j}{x_i}] = \Delta^{-1}(x)\mathbf{z}\Delta(x)$ where $\Delta(x) = \mathrm{diag}(x_1, \ldots, x_m)$.

Proof. Observe that

$$G(\Delta^{-1}(x)\mathbf{z}\Delta(x)) = G\left(\left[z_{ij}\frac{x_j}{x_i}\right]\right) = \sum_{\mathbf{k}} g_{\mathbf{k}}\mathbf{z}^{\mathbf{k}} \prod_{i=1}^m x_i^{\sum_j k_{ji} - \sum_j k_{ij}}. \tag{25}$$

Therefore, $G^*(\mathbf{z})$ is the coefficient of $G([z_{ij}\frac{x_j}{x_i}])(\cdot)$ at $x_1^0 x_2^0 \cdots x_m^0$ denoted as $\left[x_1^0 \cdots x_m^0\right]$ since $\sum_j k_{ji} - \sum_j k_{ij} = 0$ for matrices $\mathbf{k} \in \mathcal{F}$. The result follows from the Cauchy coefficient formula (cf. [76]). ∎

Now we are ready to enumerate $\mathcal{F}_n(m)$. Setting in Lemma 2 $z_{ij} = zx_i/x_j$ and using (24) we conclude that

$$F_m^*(z) = \frac{1}{(1-z)^m}[x_1^0 x_2^0 \cdots x_m^0] \prod_{i\neq j}\left[1 - z\frac{x_i}{x_j}\right]^{-1}. \tag{26}$$

Thus, by the Cauchy formula

$$|\mathcal{F}_n(m)| = [z^n]F_m^*(z) = \frac{1}{2\pi i}\oint \frac{F_m^*(z)}{z^{n+1}} dz.$$

This allows us to formulate our main result on the enumeration of (cyclic) Markov types.

Theorem 3 (Knessl, Jacquet, and Szpankowski, 2012)**.** (i) Cyclic Types. *For fixed m and $n \to \infty$ the number of cyclic Markov types is*

$$|\mathcal{P}_n(m)| = d(m)\frac{n^{m^2-m}}{(m^2-m)!} + O(n^{m^2-m-1}) \tag{27}$$

where $d(m)$ is a constant that also can be expressed by the following integral

$$d(m) = \frac{1}{(2\pi)^{m-1}} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{(m-1)-fold} \prod_{j=1}^{m-1} \frac{1}{1+\varphi_j^2} \cdot \prod_{k \neq \ell} \frac{1}{1+(\varphi_k - \varphi_\ell)^2}\, d\varphi_1 d\varphi_2 \cdots d\varphi_{m-1}. \tag{28}$$

When $m \to \infty$ we find that

$$|\mathcal{P}_n(m)| \sim \frac{\sqrt{2} m^{3m/2} e^{m^2}}{m^{2m^2} 2^m \pi^{m/2}} \cdot n^{m^2-m} \tag{29}$$

provided that $m^4 = o(n)$.
(ii) Markov Types. *The number of Markov types $|\mathcal{Q}_n(m)|$ with arbitrary initial conditions satisfies*

$$|\mathcal{Q}_n(m)| = (m^2 - m + 1)|\mathcal{P}_n(m)|(1 - O(n^{-2m}))$$

where $|\mathcal{P}_n(m)|$ is presented in (i).

In order to finish our analysis of the minimax redundancy, we need to estimate the number of sequences of a given type. First, we replace (16) by

$$D_n(\mathcal{M}_1) = m \sum_{b \in \mathcal{A}} \sum_{\mathbf{k} \in \mathcal{F}_n, k_{ba} > 0} |\mathcal{T}_n^{ba}(\mathbf{k} - [\delta_{ba}])| \mathbf{k}^{-[\delta_{ba}]} (k_b - 1)^{-k_b+1} \prod_{i \neq b} (k_i)^{-k_i}, \tag{30}$$

where $|\mathcal{T}_n^{ba}(\mathbf{k})|$ is the number of cyclic strings x^n of type $\mathbf{k}$ starting with b and ending with a. To compute it we first introduce

$$B_{\mathbf{k}} = \binom{k_1}{k_{11} \cdots k_{1m}} \cdots \binom{k_m}{k_{m1} \cdots k_{mm}}$$

where $k_i = \sum_j k_{ij}$. It may be viewed as the number of ways to depart from all m vertices in the multiple graph G associated with the frequency matrix $\mathbf{k}$ (but not necessarily completing an Eulerian cycle). Observe that

$$B(\mathbf{z}) = \sum_{\mathbf{k}} B_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} = \prod_{a \in \mathcal{A}} (1 - \sum_{b \in \mathcal{A}} z_{a,b})^{-1}.$$

Lemma 2 yields [37]

$$B^*(\mathbf{z}) = \sum_{\mathbf{k} \in \mathcal{F}_n(m)} B_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} = \frac{1}{\det(\mathbf{I} - \mathbf{z})}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix. Using this approach we can finally estimate the number of sequences starting with an a and finishing with a b of a given Markov type as follows

$$|\mathcal{T}_n^{ba}(\mathbf{k})| = \frac{k_{ba}}{k_b} B_{\mathbf{k}} \cdot \det_{bb}(\mathbf{I} - \mathbf{k}^*) \left(1 + O\left(\frac{1}{n}\right)\right),$$

where $\mathbf{k}^*$ is the matrix whose ij-th element is k_{ij}/k_i, that is, $\mathbf{k}^* = [k_{ij}/k_i]$ (cf. [83, 37]).

Putting everything together, in [37] we prove the following asymptotic expansion for the minimax redundancy of Markov sources (cf. [59]).

Theorem 4 (Rissanen, 1996, Jacquet and Szpankowski, 2004)**.** (i) *Let $\mathcal{M}_1$ be the class of Markov sources over a finite alphabet $\mathcal{A}$ of size m. The worst case minimax redundancy is $R_n^*(\mathcal{M}_1) = \log D_n(\mathcal{M}_1)$ where*

$$D_n(\mathcal{M}_1) = \left(\frac{n}{2\pi}\right)^{m(m-1)/2} A_m \times \left(1 + O\left(\frac{1}{n}\right)\right) \tag{31}$$

with

$$A_m = m \int_{\mathcal{K}(1)} F_m(y_{ij}) \prod_{i \in \mathcal{A}} \frac{\sqrt{\sum_{j \in \mathcal{A}} y_{ij}}}{\prod_{j \in \mathcal{A}} \sqrt{y_{ij}}} \prod_{ij \in \mathcal{A}^2} dy_{ij}$$

where

$$\mathcal{K}(1) = \{y_{ij} : y_{ij} \geq 0, \ \sum_{ij} y_{ij} = 1, \forall i : \sum_j y_{ij} = \sum_j y_{ji}\},$$

$F_m(\mathbf{y}) = \sum_{b \in \mathcal{A}} \det_{bb}(1 - \mathbf{y}^*)$, and $\mathbf{y}^*$ is the matrix whose ij-th coefficient is $y_{ij}/\sum_{j'} y_{ij'}$.

(ii) *Let $\mathcal{M}_r$ be the class of Markov sources of order r over a finite alphabet $\mathcal{A}$ of size m. The minimax redundancy is $R_n^*(\mathcal{M}_1) = \log D_n(\mathcal{M}_r)$ where*

$$R_n^*(\mathcal{M}_r) = \left(\frac{n}{2\pi}\right)^{m^r(m-1)/2} A_m^r \times \left(1 + O\left(\frac{1}{n}\right)\right) \tag{32}$$

with

$$A_m^r = m^r \int_{\mathcal{K}_r(1)} F_m^r(\mathbf{y}) \prod_{w \in \mathcal{A}^r} \frac{\sqrt{y_w}}{\prod_j \sqrt{y_{w,j}}},$$

where $\mathcal{K}_r(1)$ is the convex set of $m^r \times m$ matrices $\mathbf{y}$ with non-negative coefficients such that $\sum_{w,j} y_{w,j} = 1$, $w \in \mathcal{A}^r$. The function

$$F_m^r(\mathbf{y}_r) = \sum_w \det_{ww}(\mathbf{I} - \mathbf{y}_r^*),$$

where $\mathbf{y}_r^*$ is the $m^r \times m^r$ matrix whose (w, w') coefficient is equal to $y_{w,a}/\sum_{i \in \mathcal{A}} y_{wi}$ if there exist a in $\mathcal{A}$ such that w' is a suffix of wa, otherwise the (w, w')th coefficient is equal to 0.

The evaluation of the constants A_m is not easy. But, for a binary alphabet ($m = 2$) we have

$$A_2 = 2 \int_{\mathcal{K}(1)} (\det_{11}(\mathbf{I} - \mathbf{y}^*) + \det_{22}(\mathbf{I} - \mathbf{y}^*)) \frac{\sqrt{y_1}}{\sqrt{y_{11}}\sqrt{y_{12}}} \frac{\sqrt{y_2}}{\sqrt{y_{21}}\sqrt{y_{22}}} dy_{11} dy_{12} dy_{21} dy_{22}. \tag{33}$$

Since $\det_{11}(\mathbf{I} - \mathbf{y}^*) = \frac{y_{21}}{y_2}$ and $\det_{22}(\mathbf{I} - \mathbf{y}^*)$ by symmetry, and since the condition $\mathbf{y} \in \mathcal{K}(1)$ means $y_1 + y_2 = 1$ and $y_{12} = y_{21}$ we arrive at, $A_2 = 16 \cdot G$ where G is the Catalan constant defined as $G = \sum_i \frac{(-1)^i}{(2i+1)^2} \approx 0.915965594$.

3 Science of Information: Beyond Shannon

In *science of information* the goal is to pursue the theory of information beyond Shannon's original objectives (of communication), by applying it to problems of biology, neuroscience, economics, physics, and massive data where knowledge extraction is the game changer. We believe that in order to make fundamental contributions to these applications, we first need better understanding of new aspects of temporal, spatial, structural and semantic information. In this section, we first briefly review some recent results on semantic and temporal properties and on cooperation, to focus on structural information.

3.1 Delay, Semantic, and Cooperation

The mathematical theory of information arose from Shannon's theorem on channel capacity, defined as the maximum rate that can be achieved over a channel with asymptotically small probability of error. Shannon capacity of a channel places no restrictions on complexity or **delay** in transmission or reception. Methods to properly characterize the complexity and the delay could potentially fill a large gap that would extend Shannon capacity to dynamic networks with multi-point communication and often unpredictable delays [26, 29]. Furthermore, the increasing demands for using wireless networks require such delay guarantees. Applications include VoIP, video streaming, real time surveillance, networked control, etc. One common characteristic of these applications is that they have a strict deadline associated with each packet. Further, the channel reliabilities of different clients can be different, and can even vary over time. These are compelling reasons why we need to understand the role of delay in distributed communication.

In [58] Polyanskiy, Poor, and Verdu extend the fundamental channel coding theorem of Shannon to a finite blocklength regime. In particular, it is shown that coding rate $M_n^*(n, \varepsilon)$ for finite block length n is

$$\frac{1}{n} \log M^*(n, \varepsilon) \approx C - \sqrt{\frac{V}{n}} Q^{-1}(\varepsilon)$$

where C is the capacity, V is the channel dispersion, ε is error probability, and Q is the complementary Gaussian distribution. This is a non-asymptotic result (i.e., precise lower and upper bounds are presented), and it allows us to compute the degradation in capacity, even for small blocklengths. Recently, these results are extended to lossy compression [47].

In another line of research in a real time coding system with lookahead, Asnani and Weissman [2] investigate the impact of delay on expected distortion. The system consists of a memoryless source; a memoryless channel; an encoder, which encodes the source symbols sequentially, with knowledge of future source symbols up to a fixed finite lookahead, with or without feedback of the past channel output symbols; and a decoder, which sequentially constructs the source symbols using the channel output. The objective is to minimize the expected per-symbol distortion using a control theory approach. The authors provide one of the first results in this line of research. This bridges the gap between causal encoding (delay=0) and the infinite lookahead case (delay=∞) where Shannon theoretic arguments show that encoding-decoding separation is optimal.

However, any further progress in information theory of networks requires us to understand distributed information and link delay with flow of information. In a novel line of research P.R. Kumar and co-authors [31, 42] design reliable scheduling policies with delay constraints for unreliable wireless networks. They focus on a formulation that appears to provide a useful and tractable framework for modeling, analyzing and designing real-time wireless communications. This framework is built on top of an analytical model that jointly considers the three important aforementioned challenges: a strict deadline for each packet, the timely throughput requirement specified by each client or application, and finally the unreliable and heterogeneous nature of wireless transmissions. An important feature is that this model is suitable for characterizing the needs of a wide range of applications, and the model allows each application to specify its individual demand.

We turn now our attention to **semantic** aspect of information. Shannon in his 1948 paper asserted "Frequently the messages have meaning, that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem." However, Sudan and his collaborators [25, 40] argue that the meaning of information does start to become relevant whenever there is diversity in the communicating parties and when parties themselves evolve over time. For example, when a computer attempts to communicate with a printer they must talk the same language in the same format (i.e., "printer driver"). This leads Sudan and his collaborators to consider communication in the setting where encoder and decoder do not agree a priori on the communication protocols, thus encoder and decoder do not understand each other. In [25, 40] a mathematical theory of goal-oriented communication is proposed from the complexity theory point of view. Perhaps these are among the first results that may lead to a new information theory of semantic communication.

Finally, we discuss information theory of **cooperation and dependency**. In an extension of the Shannon framework, Cuff, Permuter and Cover [17] initiate a theory of cooperation and coordination in networks. A general understanding of the limits of dependence yields rate distortion theory (data compression) as a special case and provides a general approach to distributed data compression and cooperation. It also elucidates such diverse processes as intercellular biological communication. The role of dependence is exemplified by the telephone system, wireless communication, the internet, news services, the economies of large countries and the internal workings of computer architecture. The efficacy of all of these systems depends on fast communication and consequent cooperative behavior. Such distributed dependence is also found in chemical reactions, landslides, hurricanes, the dynamics of the sun and the universe itself. What are the necessary information exchanges? What limits on physical dependence are imposed by the speed of information? Are there energy constraints on computation? Some of these vast generalities can be addressed by developing a science of information for dependence. In [17] the authors ask what dependence can be established among nodes given communication constraints. More precisely, the authors compute the achievable joint distribution among network nodes, provided that the communication rates are given. Such a distributed cooperation can be the solution to many problems, such as distributed games, distributed control, and bounds on the influence of one part of a physical system on another.

Dependency and rational expectation are critical ingredients in Sims' work on modern dynamic economic theory [51]. Sims points out that existing theories of rational expectations with continuous optimization imply infinite mutual information between market and person actions. By imposing information flow constraints, discrete behavior emerges (as already seen in [61]) that better describe real economic behavior (cf. also [71]).

3.2 Information Content of Graphical Structures

Structural information appears in myriad applications, from biology to social networks to material sciences. In fact, in recent years we have become inundated with new (unconventional) data: the internet, social networks, biological networks, and medical records are all key examples that present grand challenges. For instance, in recent paper [80] Varshney *et al.* reported a pretty complete wiring (graph) of 302 neurons in the *C.elegans* worm that allows inference of biological functions from the neuronal network structure.

Unconventional data often are represented by more sophisticated data structures such as graphs, sets, and trees. For example, a graph can be described by a binary matrix that further can be viewed as a binary sequence. However, such a sequence does not exhibit internal symmetries that are conveyed by the so-called graph automorphism (such automorphisms make certain sequences/matrices "indistinguishable"). The main challenge in dealing with such structural data is to identify and describe these structural relations. In fact, these "regular properties" constitute "useful (extractable) information" discussed in Section 2.1. Furthermore, such data structures often have two types of information: the information conveyed by the structure itself, and the information conveyed by the data labels implanted in the structure. We still do not have good metrics of information embodied in structure.

As the first step in understanding structural information, we restrict our attention to structures on graphs, specifically, we study *unlabeled graphs* (or structures). In particular, given n distinguishable (labeled) vertices, a random graph is generated by adding edges randomly. This random graph model $\mathcal{G}$ produces a probability distribution on graphs, and the graph entropy $H_{\mathcal{G}}$ is defined naturally as

$$H_{\mathcal{G}} = \mathbf{E}[-\log P(G)] = -\sum_{G \in \mathcal{G}} P(G) \log P(G),$$

where $P(G)$ is the probability of a graph G. However, to focus on structural properties, we consider here unlabeled graphs in which the vertices are indistinguishable. We denote such an unlabeled graph by $S \in \mathcal{S}$ and clearly

$$P(S) = \sum_{G \cong S, G \in \mathcal{G}} P(G).$$

Here $G \cong S$ means that G and S have the same structure, that is, S is *isomorphic* to G. Thus, if all isomorphic labeled graphs have the same probability, then for any labeled graph $G \cong S$,

$$P(S) = N(S) \cdot P(G), \tag{34}$$

where $N(S)$ is the number of different labeled graphs that have the same structure as S. The *structural entropy* $H_{\mathcal{S}}$ of a random graph can be defined as the entropy of a random structure $\mathcal{S}$, that is,

$$H_{\mathcal{S}} = \mathbf{E}[-\log P(S)] = -\sum_{S \in \mathcal{S}} P(S) \log P(S),$$

where the summation is over all distinct structures.

In order to compute the probability of a given structure S, one needs to estimate the number of ways, $N(S)$, to construct a given structure S (i.e., unlabeled graph). For this, the automorphisms of a graph is to be considered. An *automorphism* of a graph G is an *adjacency preserving permutation* of the vertices of G. The collection $\mathrm{Aut}(G)$ of all automorphisms of G is called *the automorphism group* of G. In the sequel, $\mathrm{Aut}(S)$ of a structure S denotes $\mathrm{Aut}(G)$ for some labeled graph G such that $G \cong S$. In group theory, it is well known that

$$N(S) = \frac{n!}{|\mathrm{Aut}(S)|}$$

and therefore, $1 \leq |\mathrm{Aut}(S)| \leq n!$.

This trivial observation leads to a relation between the graph entropy and the structural entropy [10].

Lemma 4. *If all isomorphic graphs have the same probability, then*

$$H_{\mathcal{S}} = H_{\mathcal{G}} - \log n! + \sum_{S \in \mathcal{S}} P(S) \log |\mathrm{Aut}(S)|$$

for any random graph $\mathcal{G}$ and its corresponding random structure $\mathcal{S}$, where $\mathrm{Aut}(S)$ *is the automorphism group of S.*

In order to further advance our theory, we need to adopt a graph generation model. From now on, we assume a memoryless Erdős-Rényi model $\mathcal{G}(n,p)$ over n vertices in which edges are added independently and randomly with probability p. Thus $P(G) = p^k q^{\binom{n}{2}-k}$, where $q = 1-p$. To compute the entropy of $\mathcal{S}(n,p)$ we need to estimate $N(S)$. For this, we must study an important property of $\mathcal{G}(n,p)$, namely *asymmetry*. A graph is said to be *asymmetric* if its automorphism group does not contain any permutation other than the identity (i.e., $|\mathrm{Aut}(G)| = 1$); otherwise it is called *symmetric*. It is known that almost every graph from $\mathcal{G}(n,p)$ is asymmetric [7, 43]. In the sequel, we write $a_n \ll b_n$ to mean $a_n = o(b_n)$ when $n \to \infty$.

Lemma 5 (Kim, Sudakov, and Vu, 2002)**.** *For all p satisfying $\frac{\ln n}{n} \ll p$ and $1-p \gg \frac{\ln n}{n}$, a random graph $G \in \mathcal{G}(n,p)$ is symmetric with probability $O\left(n^{-w}\right)$ for any positive constant w.*

Using this property, we can now present the structural entropy and establish the asymptotic equipartition property (AEP), that is, the typical probability of a structure S. In [10] we prove.

Theorem 5 (Choi and Szpankowski, 2009)**.** *For large n and all p satisfying $\frac{\ln n}{n} \ll p$ and $1-p \gg \frac{\ln n}{n}$, the following holds:*
(i) *The structural entropy $H_{\mathcal{S}}$ of $\mathcal{G}(n,p)$ is*

$$H_{\mathcal{S}} = \binom{n}{2} h(p) - \log n! + O\left(\frac{\log n}{n^{\alpha}}\right), \qquad \text{for some} \ \ \alpha > 0,$$

(ii) *(AEP) For a structure $S \in \mathcal{S}(n,p)$ and $\varepsilon > 0$,*

$$P\left(\left|-\frac{1}{\binom{n}{2}}\log P(S) - h(p) + \frac{\log n!}{\binom{n}{2}}\right| < \varepsilon\right) > 1 - 2\varepsilon, \tag{35}$$

where $h(p) = -p\log p - (1-p)\log(1-p)$ is the entropy rate of a binary memoryless source.

By Shannon's source coding theorem, the structural entropy computed in Theorem 5 is a fundamental lower bound for the lossless compression of structures from $\mathcal{S}(n,p)$. However, the challenge is to design an asymptotically optimal compression algorithm matching the first two leading terms $\binom{n}{2}h(p) - n\log n$ of the structural entropy with high probability. We discuss it next.

Our algorithm, called Szip (Structural zip), is a compression scheme for unlabeled graphs. In other words, given a labeled graph G, it compresses G into a codeword, from which one can construct a graph S that is isomorphic to G. The algorithm consists of two stages. First it encodes G into two binary sequences and then compresses them using an arithmetic encoder.

The main idea behind our algorithm is quite simple (see [10] for details and Figure 7): We select a vertex of a graph, say v_1 ($v_1 = i$ in Figure 7), and store the *number* of neighbors of v_1 in a binary string B_1 (0100 in Figure 7). We remove this vertex, and then partition the remaining $n-1$ vertices into two sets: the neighbors of v_1 (d, f, g, i in Figure 7) and the non-neighbors of v_1 (a, b, c, e, h in Figure 7). We continue by selecting (and removing) a vertex, say v_2 ($v_2 = f$ in Figure 7), from the neighbors of v_1 and store two *numbers* in either string B_1 or B_2 (if there is only one neighbor or none): the number of neighbors of v_2 among each of the above two sets. Then we partition the remaining $n-2$ vertices into four sets: the neighbors of both v_1 and v_2, the neighbors of v_1 that are non-neighbors of v_2, the non-neighbors of v_1 that are neighbors of v_2, and the non-neighbors of both v_1 and v_2. This procedure continues until all vertices are processed. This process of selecting and splitting vertices can be described by a tree as illustrated in Figure 7.

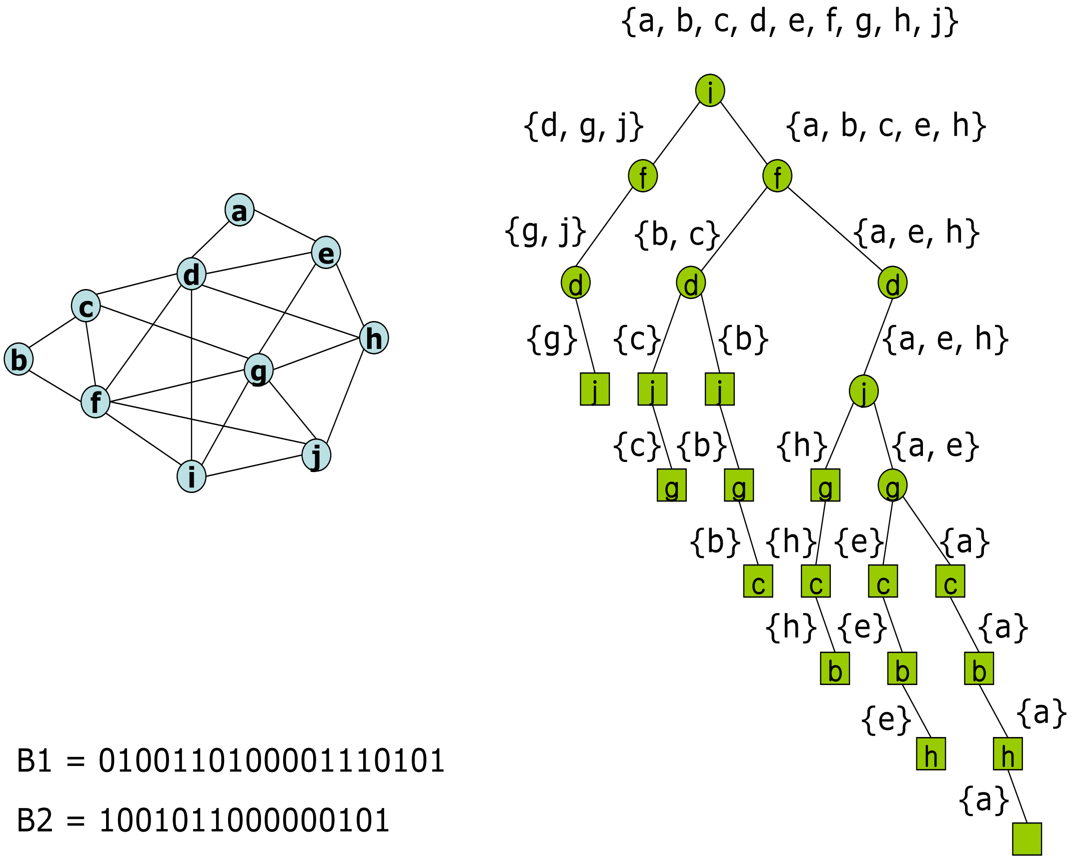


Figure 7: Illustration to Szip

During the construction the number of neighbors of the selected vertex is appended to either sequence B_1 or sequence B_2, where B_2 contains those numbers for singleton sets (i.e., we store either "0" when there is no neighbor or "1" otherwise). The sequence B_2 is represented by a "square" in the associated tree in Figure 7. We then compress B_1 and B_2 using an arithmetic encoder.

In [10] we prove that the algorithm just presented achieves the structural entropy *up to the first two leading terms* by showing that the length of B_2 (in compressed form) dominates the compression rate. In fact, we also observe that by the construction B_2 can be viewed as generated by a memoryless source with probability p. We prove the following.

Theorem 6 (Choi and Szpankowski, 2009)**.** *Let $L(S)$ be the length of the codeword generated by our algorithm for Erdős-Rényi graphs $G \in \mathcal{G}(n,p)$ isomorphic to a structure S. Then:*
(i) *For large n,*

$$\mathbf{E}[L(S)] \le \binom{n}{2} h(p) - n\log n + (c + \Phi(\log n))\, n + o(n),$$

where c is an explicitly computable constant, and $\Phi(\log n)$ is a fluctuating function with a small amplitude independent of n.

(ii) *Furthermore, for any $\varepsilon > 0$,*

$$P\left(L(S) - \mathbf{E}[L(S)] \le \varepsilon n \log n\right) \ge 1 - o(1).$$

(iii) *Our algorithm* Szip *runs either in time $O(n^2)$ in the worst case for any graph or in time $O(n+e)$ on average for graphs generated by $\mathcal{G}(n,p)$, where e is the average number of edges.*

In the remaining part of this section, we present a sketch of the proof of Theorem 6 (i). We need to compute the average lengths $L(B_1)$ and $L(B_2)$ of strings B_1 and B_2, respectively. These lengths can be evaluated through the associated tree T_n shown in Figure 7. In fact,

$$L(B_1) = \sum_{x \in T_n \text{ and } N_x > 1} \lceil \log(N_x + 1) \rceil \tag{36}$$

$$L(B_2) = \sum_{x \in T_n \text{ and } N_x = 1} \lceil \log(N_x + 1) \rceil = \sum_{x \in T_n \textit{ and } N_x = 1} 1 \tag{37}$$

where N_x is the degree of a node x in the associated tree T_n.

To analyze $L(B_1)$ and $L(B_2)$ it is convenient to introduce an auxiliary tree that we call (n,d)-tries[3] and denote as $T_{n,d}$. The root of such a tree contains n balls (vertices of the underlying graph) that are consequently distributed between two subtrees according to a simple rule: In each step, all balls independently move down to the left subtree (say with probability p) or the right subtree (with probability $1-p$), and a new node is created as long as there is at least one ball in that node. Finally, a non-negative integer d is given so that at level d or greater one ball is removed from the leftmost node before the balls move down to the next level (in our case we set $d = 0$). These steps are repeated until all balls are removed (i.e., after $n{+}d$ steps). Of interest are such tree parameters as the depth, path length (sum of all depths), size, and so forth.

We compute now the averages of $L(B_1)$ and $L(B_2)$ for a randomly generated Erdős-Rényi graph. For $L(B_1)$, in the tree $T_{n,d}$ define

$$A_{n,d} = \sum_{x \in T_{n,d} \text{ and } N_x > 1} \lceil \log(N_x + 1) \rceil,$$

and then $\mathbf{E}[L(B_1)] = a_{n,0}$. Also let $a_{n,d} = \mathbf{E}[A_{n,d}]$. Clearly, $a_{0,d} = a_{1,d} = 0$ and $a_{2,0} = 0$. For $n \ge 2$ and $d = 0$, we observe that

$$a_{n+1,0} = \lceil \log(n+1) \rceil + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (a_{k,0} + a_{n-k,k}), \tag{38}$$

$$a_{n,d} = \lceil \log(n+1) \rceil + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (a_{k,d-1} + a_{n-k,k+d-1}). \tag{39}$$

To estimate $L(B_2)$ we observe that

$$L(B_2) = \sum_{x \in T_{n,0}} N_x - B_{n,0} = \frac{n(n-1)}{2} - B_{n,0}. \tag{40}$$

[3] A trie [44, 76] is an ordered tree data structure that stores keys usually represented by strings. Tries were introduced by de la Briandais (1959) and Fredkin (1960) who also introduced the name *trie* derived from "re*trie*val".

where $B_{n,d} = \sum_{x \in T_{n,d}, N_x > 1} N_x$. The last equality follows from the fact that the sum of N_x's for all x at level ℓ in $T_{n,0}$ is equal to $n - 1 - \ell$. Let $b_{n,d} = \mathbf{E}[B_{n,d}]$. Clearly, $b_{0,d} = b_{1,d} = 0$ and $b_{2,0} = 0$. For $n \geq 2$, we observe that to $a_{n,d}$:

$$b_{n+1,0} = n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left[b_{k,0} + b_{n-k,k}\right], \quad \text{for } n \geq 2, \tag{41}$$

and

$$b_{n,d} = n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left[b_{k,d-1} + b_{n-k,k+d-1}\right], \quad \text{for } n \geq 2,\ d \geq 1. \tag{42}$$

Indeed, recurrence (41) follows from the fact that starting with $n+1$ balls in the root node, and removing one ball, we are left with n balls passing through the root node. The root contributes n since each time a ball moves down it adds 1 to the path length. Those n balls move down to the left or the right subtrees. Let us assume k balls move down to the left subtree (the other $n - k$ balls must move down to the right subtree); this occurs with probability $\binom{n}{k} p^k q^{n-k}$. At level one, one ball is removed from those k balls in the root of the left subtree. This contributes $b_{k,0}$. There will be no removal from $n - k$ balls in the right subtree until all k balls in the left subtree are removed. This contributes $b_{n-k,k}$. Similarly, for $d > 0$ we arrive at recurrence (42).

We are then faced with the reduced problem to find asymptotic solutions of two-dimensional recurrences (38)–(39) and (41)–(42). We concentrate on the latter and follow [11].

If we let $d \to \infty$ in (42) and assume that $b_{n,d}$ tends to a limit $b_{n,\infty}$, then (42) becomes

$$b_{n,\infty} = n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left[b_{k,\infty} + b_{n-k,\infty}\right] \tag{43}$$

with $b_{0,\infty} = b_{1,\infty} = 0$. This is the same as the recurrence for the mean path length in a standard *trie*, discussed above. For example, in [44, 76] it is proved that

$$b_{n,\infty} = \sum_{\ell=2}^{n} (-1)^{\ell} \binom{n}{\ell} \frac{\ell}{1 - p^{\ell} - q^{\ell}}. \tag{44}$$

The asymptotic expansion of (43) and the above as $n \to \infty$ may be obtained by a combination of singularity analysis and depoissonization arguments (see [24, 36, 76]). We obtain

$$b_{n,\infty} = \frac{1}{h} n \log n + \frac{1}{h}\left[\gamma + \frac{h_2}{2h} + \Phi(\log_p n)\right] n + o(n), \tag{45}$$

where $h := h(p)$ is the entropy, $h_2 = p \log^2 p + q \log^2 q$, γ is the Euler constant, and $\Phi(x)$ is the periodic function

$$\Phi(x) = \sum_{k=-\infty, k \neq 0}^{\infty} \Gamma\left(-\frac{2k\pi i r}{\log p}\right) e^{2k\pi r i x}, \tag{46}$$

provided that $\log p / \log q = r/s$ is rational, with r and s being integers with $\gcd(r, s) = 1$. If $\log p / \log q$ is irrational, then the term with Φ is absent from the $O(n)$ term of (45).

Let now $\tilde{b}_{n,d} = b_{n,d} - b_{n,\infty}$ measures how the path lengths in the (n, d)-trie differs from those in a regular trie. From (42) and (43), we then obtain

$$\tilde{b}_{n,d} = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left[\tilde{b}(k, d-1) + \tilde{b}(n-k, k+d-1)\right], \quad \text{for } n \geq 2,\ d \geq 1, \tag{47}$$

which unlike (42) is a homogeneous recurrence. It turns out that the second term under the sum is negligible, which even further simplifies the recurrence. Then analytic techniques such as Mellin transform and depoissonization can be applied leading to asymptotic solution of (47).

We summarize our main result proved in [11].

Theorem 7. *For $n \to \infty$ and $d = O(1)$ we have $\tilde{b}(n,d) = O(\log^2 n)$. More precisely*

$$\tilde{b}_{n,d} = \frac{1}{2h\log p}\log^2 n + \frac{d}{h}\log n + \left[-\frac{1}{2h} + \frac{1}{h\log p}\left(\gamma + 1 + \frac{h_2}{2h} + \Psi(\log_p n)\right)\right]\log n + O(1), \quad (48)$$

where $\Psi(\cdot)$ is the periodic function

$$\Psi(x) = \sum_{k=-\infty, k\neq 0}^{\infty}\left[1 + \frac{2k\pi ir}{\log p}\right]\Gamma\left(-\frac{2k\pi ir}{\log p}\right)e^{2k\pi irx} \quad (49)$$

and $\log p/\log q = r/t$ is rational, as in (46). If $\log p/\log q$ is irrational, the term involving Ψ in (48) is absent. Thus

$$b_{n,0} = \frac{1}{h}n\log n + \frac{1}{h}\left[\gamma + \frac{h_2}{2h} + \Phi(\log_p n)\right]n + O(\log^2 n)$$

for large n.

To complete the proof of Theorem 6 we need to evaluate the $a_{n,0} = \mathbf{E}[L(B_1)]$ that satisfies the set of recurrences (38)-(39). Using the same approach as above we prove in [10, 11]

$$\mathbf{E}[L(B_1)] = \frac{n}{h}A_*(-1) + o(n), \qquad A_*(-1) = \sum_{\ell=2}^{\infty}\frac{\lceil\log(\ell+1)\rceil}{\ell(\ell-1)}$$

if $\log p/\log q$ is irrational. If $\log p/\log q = r/s$ is rational, the constant $A_*(-1)$ must be replaced by the oscillatory function

$$\sum_{k=-\infty, k\neq 0}^{\infty} A_*\left(-1 + \frac{2k\pi ir}{\log p}\right)e^{2k\pi ir\log_p n} \quad (50)$$

where

$$A_*(s) = \sum_{n\geq 2}\frac{\lceil\log(n+1)\rceil}{n!}\Gamma(n+s).$$

Summing up, we compute $\mathbf{E}[L(S)] = \mathbf{E}[L(\hat{B}_1) + L(\hat{B}_2)] + O(\log n)$, where $\hat{B}_1$ and $\hat{B}_2$ are strings B_1 and B_2 compressed by the arithmetic encoder, while $O(\log n)$ bits are needed to encode n. The arithmetic encoder can compress a binary sequence of length m on average up to $mh + \frac{1}{2}\log m + O(1) = mh + O(\log m)$, where h is the entropy rate of the binary source. For string B_2 we know that $h = h(p)$, and this completes the part (i) of Theorem 6(i).

Acknowledgment

This work is dedicated to Philippe Flajolet who passed away in March 2011. Philippe was a friend, a colleague, and a mentor who taught many of us analytic combinatorics. I am grateful to my other “French connection”, Philippe Jacquet, a friend of the last thirtysomething years with whom I wrote many long papers and whose cheerful attitude towards life makes research fun. I have benefited a lot from my "Palo Alto" connections: Tom Cover, whose untimely death occurred during the preparation of this paper, was a kind and great host during my sabbatical at Stanford in 1999. I thank Gadiel Seroussi and Marcelo Weinberger for many hours of discussion in HPL. Yiannis Kontoyiannis and Mark Ward kindly agreed to read this paper and complained so a better, gentle and more readable paper is presented to the readers. I am grateful to Sergio Verdu for many comments regarding my work, but mostly for his friendship and setting high standards for all of us. Finally, I thank all my co-authors for patience.

References

[1] V. Anantharam, G. Caire, M. Costa, G. Kramer, R. Yeung and S. Verdu, New Perspectives on Information Theory, *Information Theory Society Newsletter*, 62, 21-27, 2012.

[2] H. Asnani and T. Weissman, On Real Time Coding with Limited Lookahead, *Allerton Conference*, 2011.

[3] V. Balasubramanian. Statistical inference, occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, 9:349–368, 1997.

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

[5] C. Bennett, Logical Depth and Physical Complexity, in *The Universal Turing Machine: A Half Century Survey*, 1988.

[6] P. Billingsley, Statistical Methods in Markov Chains, *Ann. Math. Statistics*, 32, 12-40, 1961.

[7] B. Bollobas, *Random Graphs*, Cambridge University Press, Cambridge, 2001.

[8] F. Brooks. Three great challenges for half-century-old computer science. *J. the ACM*, 50:25–26, 2003.

[9] C. Brukner and A. Zeilinger. Conceptual inadequacy of the shannon information in quantum measurements. *Phys. Rev.*, A 63:3354–3360, 2001.

[10] Y. Choi and W. Szpankowski. Compression of graphical structures: Fundamental Limits, Algorithms, and Experiments, *IEEE Trans. Information Theory*, 55, 2, 620-638, 2012.

[11] Y. Choi, C. Knessl and W. Szpankowski, On a Recurrence Arising in Graph Compression, *23rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, AofA'12, Montreal, 2012.

[12] R. Corless, G. Gonnet, D. Hare, D. Jeffrey and D. Knuth, "On the Lambert W Function," *Adv. Computational Mathematics*, 5, pp. 329–359, 1996.

[13] T. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[14] I. Csziszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[15] I. Csiszàr and P. Shields, Redundancy Rates for Renewal and Other Processes, *IEEE Trans. Information Theory*, 42, 2065–2072, 1996.

[16] I. Csziszár, The Method of Types, *IEEE Trans. Information Theory*, 44, 2505-2523, 1998.

[17] P. Cuff, H. Permuter, and T. Cover, Coordination Capacity, *IEEE Trans. on Info. Theory*, 55, 2010.

[18] L. D. Davisson, Universal Noiseless Coding, *IEEE Trans. Information Theory*, 19, 783-795, 1973.

[19] L. D. Davisson, Minimax Noiseless Universal coding for Markov Sources, *IEEE Trans. Information Theory*, 29, 211 - 215, 1983.

[20] A. Dimitrov, A. Lazar, and J. Victor (Eds.), Methods of Information Theory, Special Issue of *Journal of Computational Neuroscience*, 30, 1, 2011.

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

[22] M. Drmota, Y. Reznik, and W. Szpankowski, Tunstall Code, Khodak Variations, and Random Walks *IEEE Trans. Information Theory*, 56, 2928 - 2937, 2010.

[23] P. Flajolet and W. Szpankowski, "Analytic Variations on Redundancy Rates of Renewal Processes," *IEEE Trans. Information Theory*, 48, pp. 2911–2921, 2002.

[24] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*, Cambridge University Press, Cambridge, 2008.

[25] O. Goldreich, B. Juba, and M. Sudan, A Theory of Goal-Oriented Communication. *Electronic Colloquium on Computational Complexity*, 2009.

[26] A. Goldsmith, M. Effors, R. Koetter, M. Medard, and L. Zheng, Beyond Shannon: The Quest for Fundamental Performance Limits of Wireless Ad Hoc Networks, *IEEE Communications Magazine*, 195-205, May 2011

[27] L. Goodman, Exact Probabilities and Asymptotic Relationships for Some Statistics from m-th Order Markov Chains, *Annals of Mathematical Statistics*, 29, 476-490, 1958.

[28] P. Grunwald. *The Minimum Description Length Principle.* The MIT Press, Cambridge, 2007.

[29] B. Hajek and A. Ephremides. Information theory and communication networks: An unconsummated union. *Trans. Information Theory*, 44:2416–2434, 1998.

[30] G. Han and B. Marcus, Asymptotics of the input-constrained binary symmetric channel capacity, *Annals of Applied Probability*, 19, 1063-1091, 2009.

[31] I. Hou and P.R. Kumar, Queueing Systems with Hard Delay Constraints: A Framework and Solutions for Real-Time Communication over Unreliable Wireless Channels, *Queueing Systems: Theory and Applications.*

[32] R. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2:22–26, 1996.

[33] P. Jacquet, B. Mans, and G. Rodolakis. Information propagation speed in delay tolernat networks: Analytic upper bounds. *Tans. Information Theory*, 56, 5001-5015 , 2010.

[34] P. Jacquet, C. Knessl and W. Szpankowski, Counting Markov Types, Balanced Matrices, and Eulerian Graphs, *Tans. Information Theory*, 58, 2012.

[35] P. Jacquet, and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161–197, 1995.

[36] P. Jacquet, and W. Szpankowski, Analytical Depoissonization and Its Applications, *Theoretical Computer Science*, 201, 1–62, 1998.

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

[38] P. Jacquet and W. Szpankowski, Noisy Constrained Capacity for BSC, *IEEE Trans. Information Theory*, 56, 5412- 5423, 2010.

[39] P. Jacquet and W. Szpankowski, Limiting Distribution of Lempel Ziv'78 Redundancy, *2011 International Symposium on Information Theory*, 1609-1612, St. Petersburg, 2011.

[40] B. Juba and M. Sudan. Universal semantic communication. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, 2008.

[41] L. E. Kay. *Who Wrote the Book of Life*. Stanford University Press, Stanford, 2000.

[42] K.D. Kim and P.R. Kumar, A Real-Time Middleware for Networked Control Systems and Application to an Unstable System, *IEEE Transactions on Control Systems Technology.*

[43] J.H. Kim, B. Sudakov, and V.H. Vu, On the asymmetry of random regular graphs and random graphs, *Random Structures and Algorithms*, 21(3-4), 216–224, 2002.

[44] D. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3, Second Edition, Addison-Wesley, Reading, MA, 1998.

[45] J. Konorski and W. Szpankowski. What is information? *Festschrift in Honor of Jorma Rissanen*, pages 154–172, 2008.

[46] I. Kontoyiannis, An Implementable Lossy Version of the Lempel-Ziv Algorithm — Part I: Optimality for Memoryless Sources, *IEEE Trans. Information Theory*, 45, 2285–2292, 1999.

[47] V. Kostina and S. Verdu, Fixed-Length Lossy Compression in the Finite Blocklength Regime: Discrete Memoryless Sources, *2010 IEEE Int. Symposium on Information Theory*, St. Petersburg, 2011.

[48] S. Lonardi, W. Szpankowski, and M. Ward, Error Resilient LZ'77 Data Compression: Algorithms, Analysis, and Experiments, *IEEE Trans. Information Theory*, 53, 1799-1813, 2007.

[49] G. Louchard, and W. Szpankowski, On the Average Redundancy Rate of the Lempel-Ziv Code, *IEEE Trans. Information Theory*, 43, 2–8, 1997.

[50] T. Luczak, and W. Szpankowski, A Suboptimal Lossy Data Compression Based in Approximate Pattern Matching, *IEEE Trans. Information Theory*, 43, 1439–1451, 1997.

[51] F. Matejka and C. Sims, Discrete Actions in Information-Constrained Tracking Problems, preprint 2011.

[52] A. Martín, G. Seroussi, and M. J. Weinberger, Type classes of tree models, *Proc. ISIT 2007*, Nice, France, 2007.

[53] O. Milenkovic, G. Alterovitz, G. Battail, T. Coleman, J. Hagenauer, S. Meyn, W. Szpankowski, Information Theory in Molecular Biology and Neuroscience, *IEEE Transactions on Information Theory*, 56, 2, 2010.

[54] I Nemenman, W Bialek, and R de Ruyter van Steveninck. Entropy and information in neural spike trains: Progress on the sam pling problem. *Phys Rev E*, 69, 2004.

[55] P. Nurse. Life, logic, and information. *Nature*, 454:424–426, 2008.

[56] A. Orlitsky and N. Santhanam, "Speaking of Infinity," *IEEE Trans. Information Theory*, 50, 2215–2230, 2004.

[57] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal Compression of Memoryless Sources over Unknown Alphabets," *IEEE Trans. Information Theory*, 50, 1469–1481, 2004.

[58] Y. Polyanskiy, H. V. Poor and S. Verdu, Channel Coding Rate in the Finite Blocklength Regime, *IEEE Trans. Inf. Theory*, 56, 5, 2307-2359, 2010.

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

[60] J. Rissanen, *Information and Complexity in Statistical Modeling*, Springer, 2007.

[61] K. Rose, A mapping approach to rate-distortion computation and analysis, *IEEE Trans. Inf. Theory*, 40, 6, 1939-1952, 1994.

[62] S. Savari, Redundancy of the Lempel-Ziv Incremental Parsing Rule, *IEEE Trans. Information Theory*, 43, 9–21, 1997.

[63] G. Seroussi, On Universal Types, *IEEE Trans. Information Theory*, 52, 171-189, 2006.

[64] P. Jacquet, G. Seroussi and W. Szpankowski, On the Entropy of a Hidden Markov Process, *Theoretical Computer Science*, 395, 203-219, 2008.

[65] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, (27):379–423 623–656, 1948.

[66] C. Shannon. The lattice theory of information. *IEEE Transaction on Information Theory*, 1:105–107, 1953.

[67] C. Shannon. The Bandwagon. *IEEE Transaction on Information Theory*, 2, 3-3, 1956.

[68] P. Shields, Universal Redundancy Rates Do Not Exist, *IEEE Trans. Information Theory*, 39, 520-524, 1993.

[69] Y. Shtarkov, "Universal Sequential Coding of Single Messages," *Problems of Information Transmission*, 23, pp. 175–186, 1987.

[70] R. Stanley, *Enumerative Combinatorics*, Vol. II, Cambridge University Press, Cambridge, 1999.

[71] J. Stiglitz. The contributions of the economics of information to twentieth century economics. *Quarterly Journal of Economics*, 115:1441–1478, 2000.

[72] W. Szpankowski, Asymptotic Properties of Data Compression and Suffix Trees, *IEEE Trans. Information Theory*, 39, 1647–1659, 1993.

[73] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, 22, 1176–1198, 1993.

[74] W. Szpankowski, On Asymptotics of Certain Recurrences Arising in Universal Coding, *Problems of Information Transmission*, 34, 55-61, 1998.

[75] W. Szpankowski, Asymptotic Average Redundancy of Huffman (and Other) Block Codes, *IEEE Trans. Information Theory*, 46, 2434-2443, 2000.

[76] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, New York, 2001.

[77] W. Szpankowski and S. Verdu, Minimum Expected Length of Fixed-to-Variable Lossless Compression without Prefix Constraints *IEEE Trans. Information Theory*, 57, 4017 - 4025, 2011.

[78] W. Szpankowski and M. Weinberger, Minimax Pointwise Redundancy for Memoryless Models over Large Alphabets, *IEEE Trans. Information Theory*, 58, 2012.

[79] G. Tkacik, C. Callan, and W Bialek. Information flow and optimization in transcriptional regulation. *PNAS*, 105:12265

[80] LR. Varshney, BL. Chen, E. Paniagua DH. Hall, DB. Chklovskii, Structural Properties of the Caenorhabditis elegans Neuronal Network, *PLoS Comput Biol* 7(2), 1-21, 2011.

[81] Q. Xie, A. Barron, Minimax Redundancy for the Class of Memoryless Sources, *IEEE Trans. Information Theory*, 43, 647-657, 1997.

[82] Q. Xie, A. Barron, Asymptotic Minimax Regret for Data Compression, Gambling, and Prediction, *IEEE Trans. Information Theory*, 46, 431-445, 2000.

[83] P. Whittle, Some Distribution and Moment Formulæ for Markov Chain, *J. Roy. Stat. Soc.,* Ser. B., 17, 235-242, 1955.

[84] E.H. Yang, and J. Kieffer, On the Performance of Data Compression Algorithms Based upon String Matching, *IEEE Trans. Information Theory*, 44, 47–65, 1998.

[85] A. Zeilinger, The message of the quantum, *Nature*, 438, 743-744, 2005.