Purdue Research and Scholarship Distinction Nomination for:

W. Szpankowski
Department of Computer Science
Purdue University
USA

September 1, 2015

Outline

1. Szpankowski's Nomination Overview
2. Szpankowski's Technical Contributions
   - Entropy of Hidden Markov Process
   - Constrained Channel Capacity
   - Structural Information and Graph Compression
   - *Analytic Pattern Matching: From DNA to Twitter*, Cambridge, 2015.
3. Szpankowski's Vision: Science of Information
   - Post Shannon Information Theory
   - NSF Science and Technology Center
4. Quotes from Letter Writers

Szpankowski's Nomination Overview

Purdue Research & Scholarship Distinction Nomination of Szpankowski for:

- Solving long-standing open problems: the entropy of hidden Markov processes and the noisy constrained capacity.
- For developing innovative analytic methods for Shannon information theory leading to solutions of several open problems (e.g., Ziv's conjecture, Wyner-Ziv conjecture, Steinberg-Gutman conjecture, Huffman's code redundancy, Csiszár-Shields renewal process redundancy)
- Leading the effort of understanding structural information and designing the first asymptotically optimal graph compression algorithm.
- Visionary ideas that led first to the creation of the field of "analytic information theory," and subsequently to broadening of Shannon Information Theory to a new science of information, leading to the establishment of Indiana's first NSF Science of Technology Center for Science of Information (CSoI) and one of only two centers ever awarded in computing disciplines.
- Opening new venues of pattern matching analysis (e.g., significance analysis of patterns detection) as presented in Szpankowski's recent monograph: "*Analytic Pattern Matching: From DNA to Twitter*", Cambridge, July 2015.

Szpankowski's Nomination Overview

This nomination is based on the following recent technical results:

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

(2) P. Jacquet and W. Szpankowski, Noisy Constrained Capacity for BSC, *IEEE Transaction on Information Theory*, 56, 5412- 5423, 2010.

(3) Y. Choi and W. Szpankowski, Compression of Graphical Structures: Fundamental Limits, Algorithms, and Experiments, IEEE Trans. Information Theory, 58, 620-638, 2012

(4) M. Drmota and W. Szpankowski, A Discrete Divide and Conquer Recurrence, *Journal of the ACM*, 60, 3, 16:1-16:49, 2013.

(5) P. Jacquet and W. Szpankowski, *Analytic Pattern Matching: From DNA to Twitter*, Cambridge, 2015.

Outstanding Challenges in Computing

The most pressing challenge of our times is the data deluge and the transformation from data to information, and subsequently to knowledge.

1. 25.21 billion web pages (2009), over 1 trillion distinct URLs (2008).
2. The amount of data in the deep web far exceeds this.
3. About 56% of the text data is in English.
4. Easy Questions: How much unique data? How much information in text? Translating this information into actionable form?
5. Increasingly data is not in the form of text – social networks, tweets, scientific data (interactions, geometries, time series), economic transactions, etc.
6. Harder Questions: How do we quantify this data, how do we extract information from these datasets? How do we act on this information?
7. Really Hard Questions: Information has cause and consequence – How do we reach beyond information?

Outstanding Challenges in Computing

These are profound questions and Wojtek is an acknowledged world leader in quantitative methods addressing these problems.

- The best researchers from the premier institutions, worldwide, have rallied around him to define and promote the area (**Visibility**).

- Wojtek has solved some of the longest standing problems in the area (**Depth**).

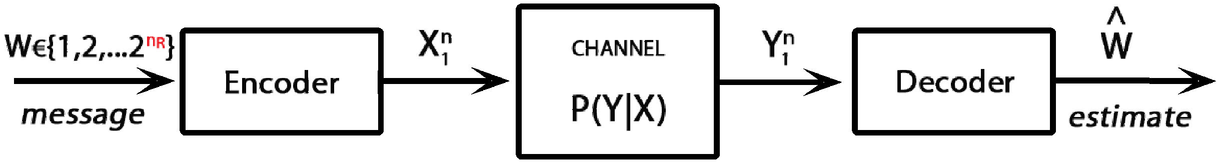
- Wojtek's unique contributions transcend computing – reaching out to scientific disciplines such as life sciences and physics (**Breadth and Impact**).

- He has driven the research agenda of the community at large, through his tireless contributions in the form of conferences, journals, and workshops (**Service**).

Three Theorems of Shannon

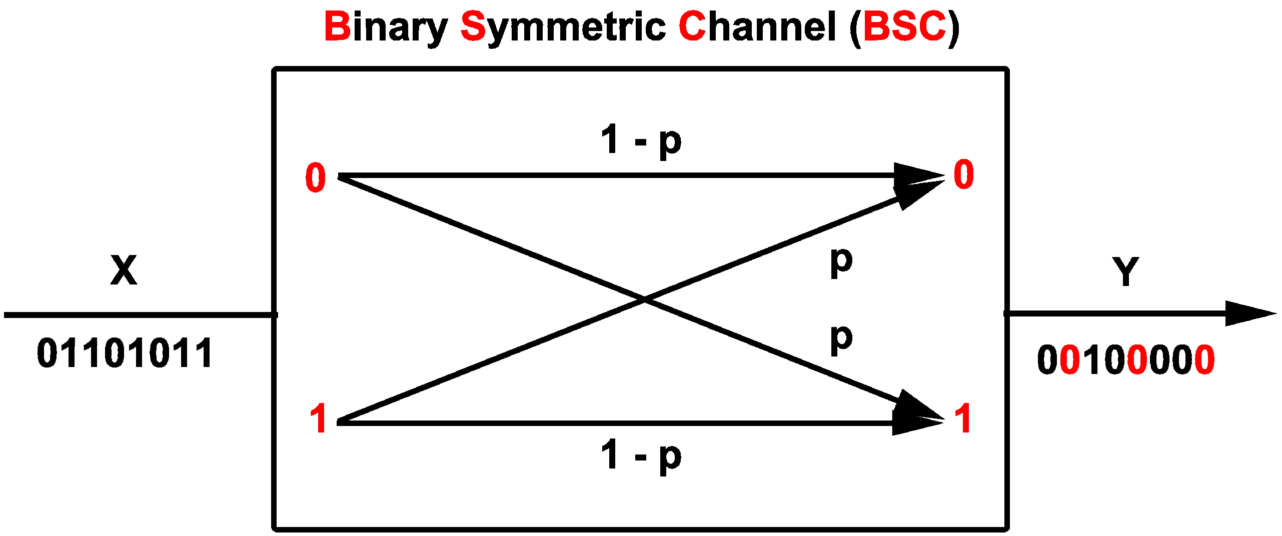
Theorem 1. (**Shannon 1948; Source Coding**) It is impossible to compress data such that the code rate (average number of bits per symbol) is less than the Shannon entropy of the source, without it being virtually certain that information will be lost. However it is possible to get the code rate arbitrarily close to the Shannon entropy, with negligible probability of loss.

Theorem 2. (**Shannon 1948; Channel Coding**)
It is possible to send information at the capacity through the channel. with as small a frequency of errors as desired by proper (**long**) encoding. This statement is not true for any rate greater than the capacity.



Theorem 3. (**Shannon 1948; Lossy Source Coding**) The minimum rate that we can represent the source X with average distortion D is given by the rate distortion function $R(D)$.

Capacity of BSC



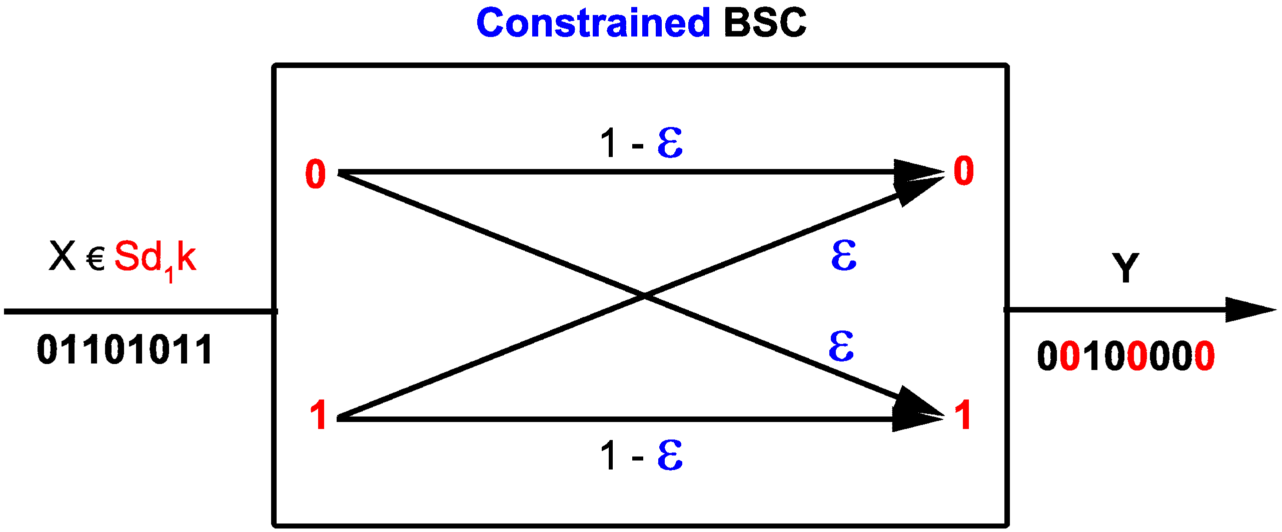
Capacity:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(p) \\ &\leq 1 - H(p). \end{aligned}$$

The capacity is achieved for the uniform input distribution. Thus

$$C = 1 - H(p).$$

(Szpankowski's 2010) Noisy Constrained Channel



Things get interesting if the input sequences are constrained. Let $\mathcal{S}$ denote the set of binary constrained sequences of length n. Here:

$$\mathcal{S}_{d,k} = \{\text{(d,k) sequences}\},$$

i.e., no sequence contains a run of zeros shorter than d or longer than k (applications: DVD, CD, blue-rays, biology).

Sequence $X \in \mathcal{S}_{(d,k)}$ can be represented as a MARKOV PROCESS.

$C(\mathcal{S}, \varepsilon)$ – noisy constrained capacity defined as

$$C(\mathcal{S}, \varepsilon) = \sup_{X \in \mathcal{S}} I(X;Y) = \lim_{n \to \infty} \frac{1}{n} \sup_{X_1^n \in \mathcal{S}_n} I(X_1^n, Y_1^n).$$

This is/was an open problem since 1948 Shannon work.

Entropy of Hidden Markov Process

Hidden Markov Process: Since

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\varepsilon)$$

$(H(\varepsilon) = -\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon))$ we need to find $H(Y)$.
But Y is a Hidden Markov Process (HMP) since it is a noisy version of the Markov Process X.

Entropy of HMP was first investigated by Blackwell in 1956 but no significant progress since then. Why?
Theorem 1 (Jacquet, Seroussi, & Szpankowski, 2004)**.** *Consider the HMP Y as defined above. The entropy rate*

$$H(Y) \quad = \quad \lim_{n\to\infty} \frac{1}{n} \mathbf{E}[-\log\left(\mathbf{p}_1 \mathbf{M}(Y_1,Y_2)\cdots\mathbf{M}(Y_{n-1},Y_n)\mathbf{1}^t\right)] = \mu(P)$$

where $\mu(P)$ is a top Lyapunov exponent of random matrices $\mathbf{M}(Y_1,Y_2)\cdots\mathbf{M}(Y_{n-1},Y_n)$ defined as

$$\mathbf{M}(Y_{n-1},Y_n) = \left[\begin{array}{cc} (1-\varepsilon)P_X(Y_n|Y_{n-1}) & \varepsilon P_X(\bar{Y}_n|Y_{n-1}) \\ (1-\varepsilon)P_X(Y_n|\bar{Y}_{n-1}) & \varepsilon P_X(\bar{Y}_n|\bar{Y}_{n-1}) \end{array}\right].$$

Asymptotic Expansion & Capacity

We now assume that $P(\mathrm{error}) = \varepsilon \to 0$ is small (never studied before).

Theorem 2 (Jacquet and Szpankowski, 2004, 2007). *Assume rth order Markov. Then the entropy rate of Y for small ε is*

$$H(Y) = H(X) - f_0(P)\varepsilon \log \varepsilon + f_1(P)\varepsilon + o(\varepsilon)$$

for explicitly computable $f_0(P)$ and $f_1(P)$.

The main result that solved a 50 year open problem is presented next.

Theorem 3 (Jacquet & Szpankowski, 2008). [1] *The capacity of the noisy constrained channel is*

$$C(\mathcal{S}, \varepsilon) = C(\mathcal{S}) - (1 - f_0(P^{\max}))\varepsilon \log \varepsilon + (f_1(P^{\max}) - 1)\varepsilon + o(\varepsilon)$$

where $C(\mathcal{S})$ is the capacity of noiseless system ($\varepsilon = 0$)

[1] In 2004 Marcus at al. stated: "... *while calculation of the noise-free capacity of constrained sequences is well known, the computation of the capacity of a constraint in the presence of noise* ... *has been an unsolved problem in the half-century since Shannon's landmark paper* ...."

Divide and Conquer

Divide and Conquer:
A divide and conquer algorithm splits the input into several smaller subproblems, solving each subproblem separately, and then knitting together to solve the original problem.

Complexity:
A problem of size n is divided into $m \geq 2$ subproblems of size $\lfloor p_j n + \delta_j \rfloor$ and $\lceil p_j n + \delta'_j \rceil$ and each subproblem contributes b_j, b'_j fraction to the final solution; there is a cost a_n associated with combining subproblems.

Total Cost:
The total cost $T(n)$ satisfies the **discrete divide and conquer** recurrence:

$$T(n) = a_n + \sum_{j=1}^{m} b_j T\left(\lfloor p_j n + \delta_j \rfloor\right) + \sum_{j=1}^{m} b'_j T\left(\left\lceil p_j n + \delta'_j \right\rceil\right) \qquad (n \geq 2)$$

where $0 \leq p_j < 1$ (e.g., $\sum_{i=1}^{m} p_i = 1$).

How to solve precisely this *discrete divide & conquer* recurrence?

Main Result – Precise Master Theorem

Theorem 4 (M. Drmota and W. Szpankowski, 2013).[2] *Let* $a_n = Cn^{\sigma a}(\log n)^{\alpha}$ *with* $\min\{\sigma, \alpha\} \geq 0$.

(i) If $\log(1/p_1), \ldots, \log(1/p_m)$ *are irrationally related, then*

$$T(n) = \begin{cases} C_1 + o(1) & \text{if } \sigma_a \leq 0 \text{ and } s_0 < 0, \\ C_2 \log n + C_2' + o(1) & \text{if } \sigma_a < s_0 = 0, \\ C_3(\log n)^{\alpha+1}(1 + +o(1)) & \text{if } \sigma_a = s_0 = 0 \\ C_4\, n^{s_0} \cdot (1 + o(1)) & \text{if } \sigma_a < s_0 \text{ and } s_0 > 0, \\ C_5 n^{s_0}(\log n)^{\alpha+1} \cdot (1 + o(1)) & \text{if } \sigma_a = s_0 > 0 \text{ and } \alpha \neq -1, \\ C_5 n^{s_0} \log\log n \cdot (1 + o(1)) & \text{if } \sigma_a = s_0 > 0 \text{ and } \alpha = -1, \\ C_6(\log n)^{\alpha}(1 + o(1)) & \text{if } \sigma_a = 0 \text{ and } s_0 < 0, \\ C_7 n^{\sigma a}(\log n)^{\alpha} \cdot (1 + o(1)) & \text{if } \sigma_a > s_0 \text{ and } \sigma_a > 0. \end{cases}$$

(ii) *If* $\log(1/p_1), \ldots, \log(1/p_m)$ *are rationally related, then* $T(n)$ *behaves as in the irrationally related case with the following two exceptions:*

$$T(n) = \begin{cases} C_2 \log n + \Psi_2(\log n) + o(1) & \text{if } \sigma_a < s_0 = 0, \\ \Psi_4(\log n)\, n^{s_0} \cdot (1 + o(1)) & \text{if } \sigma_a < s_0 \text{ and } s_0 > 0, \end{cases}$$

where C_2 *is positive and* $\Psi_2(t), \Psi_4(t)$ *are periodic functions with period* L *(with usually countably many discontinuities).*

[2] D. E. Knuth, Stanford, wrote in a postcard to W. Szpankowski "Bravo for your recent masterful paper in JACM!".

Post-Shannon Challenges

Classical Information Theory needs a recharge to meet new challenges of emerging applications in biology, modern communication, knowledge extraction, economics and physics, ...

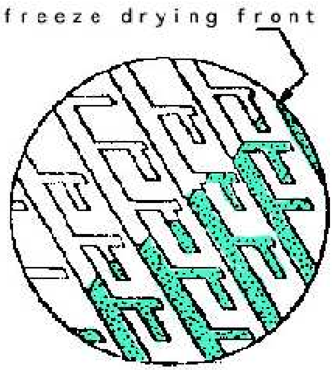
We need to extend traditional formalisms for information to include

structure, time, space, and semantics ,

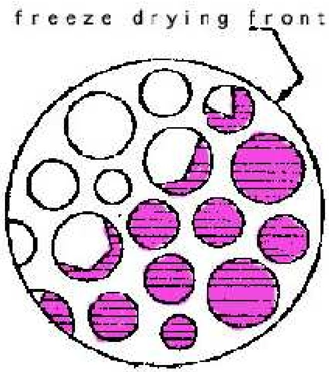
and others such as:

dynamic information, limited resources, complexity, physical information, representation-invariant information, and cooperation & dependency.

Structure:
Measures are needed for quantifying information embodied in structures (e.g., material structures, nanostructures, biomolecules, gene regulatory networks protein interaction networks, social networks, financial transactions).
(F. Brooks, JACM, 2003.)

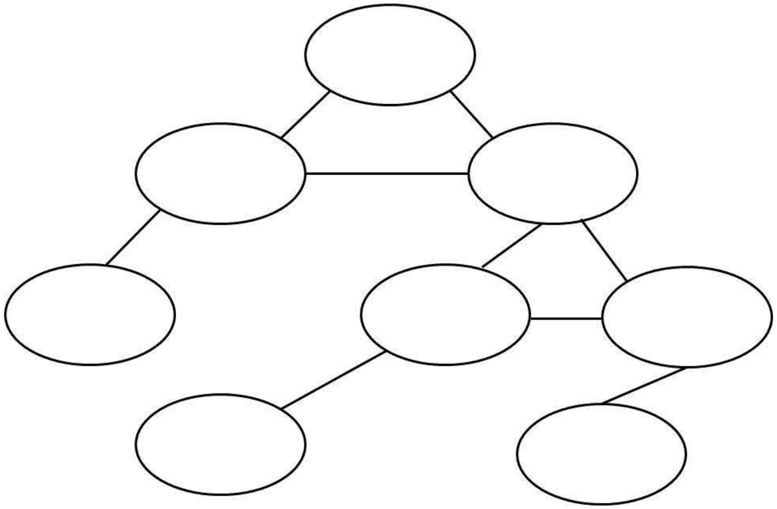


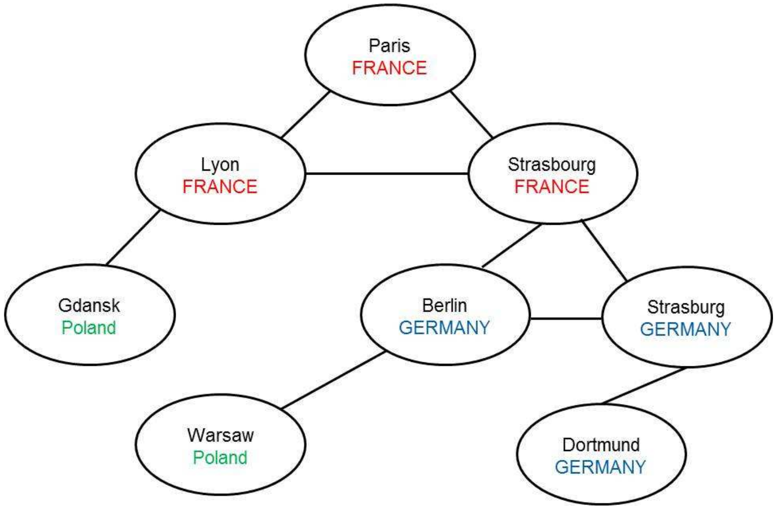
crystalline



amorphous

Information in Networks





Paris
FRANCE

Lyon
FRANCE

Strasbourg
FRANCE

Gdansk
Poland

Berlin
GERMANY

Strasburg
GERMANY

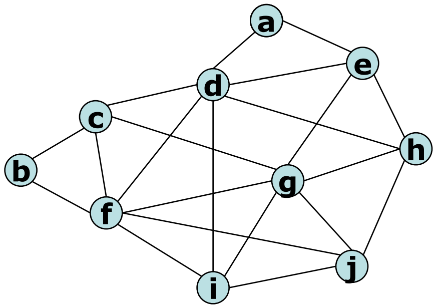
Warsaw
Poland

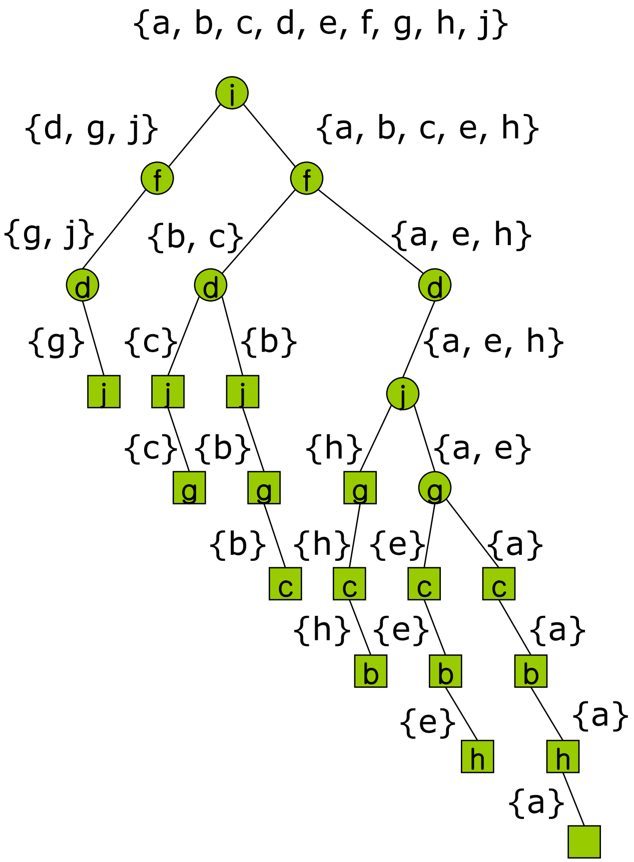
Dortmund
GERMANY

How many bits are required to describe the unlabeled graph on the left, and how many additional bits one needs to represent the correlated labels on the right?

Optimal Compression: Structural Zip (SZIP) Algorithm

Compression Algorithm called Structural zip, in short SZIP –





B1 = 0100110100001110101

B2 = 1001011000000101

Asymptotic Optimality of SZIP for Erdös-Rényi Graphs

Theorem 5 (Y. Choi and W. Szpankowski, 2012). *Let* $L(S) = |\tilde{B}_1| + |\tilde{B}_2|$ *be the code length.*

(i) *For large* n*,*

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

where c *is an explicitly computable constant, and* $\Phi(x)$ *is a fluctuating function with a small amplitude or zero.*

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

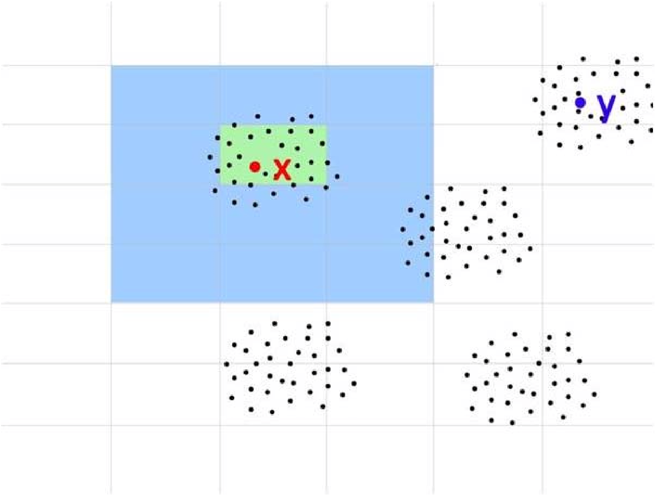
$$P\,(L(S) - \mathbf{E}[L(S)] \leq \varepsilon n \log n) \geq 1 - o(1).$$

(iii) *The algorithm runs in* $O(n + e)$ *on average, where* e # *edges.*

Table 1: The length of encodings (in bits)

	Networks	# of nodes	# of edges	our algorithm	adjacency matrix	adjacency list	arithmetic coding
Real-world	US Airports	332	2,126	8,118	54,946	38,268	12,991
	Protein interaction (Yeast)	2,361	6,646	46,912	2,785,980	1 59,504	67,488
	Collaboration (Geometry)	6,167	21,535	115,365	19,012, 861	55 9,910	241,811
	Collaboration (Erdös)	6,935	11,857	62,617	24,043,645	308,2 82	147,377
	Genetic interaction (Human)	8,605	26,066	221,199	37,0 18,710	729,848	310,569
	Internet (AS level)	25,881	52,407	301,148	334,900,140	1,572, 210	396,060

Challenges: Time & Space, and Semantics



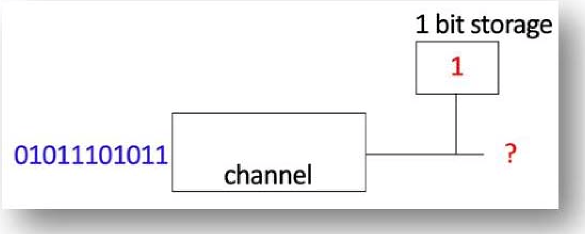
Time & Space:
Classical Information Theory is at its weakest in dealing with problems of delay (e.g., information arriving late maybe useless or has less value).
(P. Jacquet et al., IT 2010.)

Semantics & Learnable information:
Data driven science focuses on extracting information from data. How much information can actually be extracted from a given data repository? How much knowledge is in Google's database?
(M. Sudan et al., 2010.)

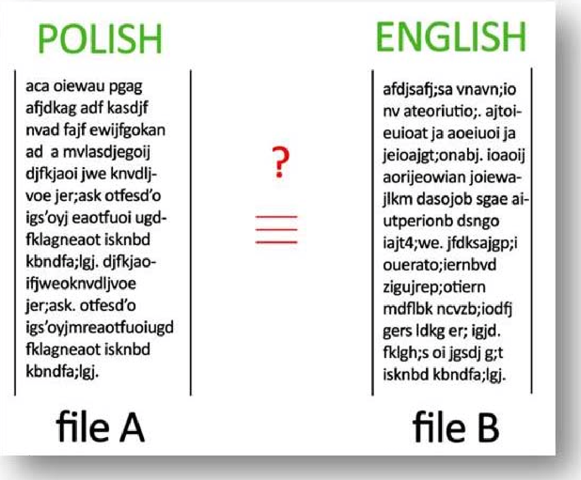
Cooperation. Often subsystems may be in conflict (e.g., denial of service) or in collusion (e.g., price fixing). How does cooperation impact information? (In wireless networks nodes should cooperate in their own self-interest.)
(Cuff, et al. IT, 2010).

Challenges: Limited Resources and Representation

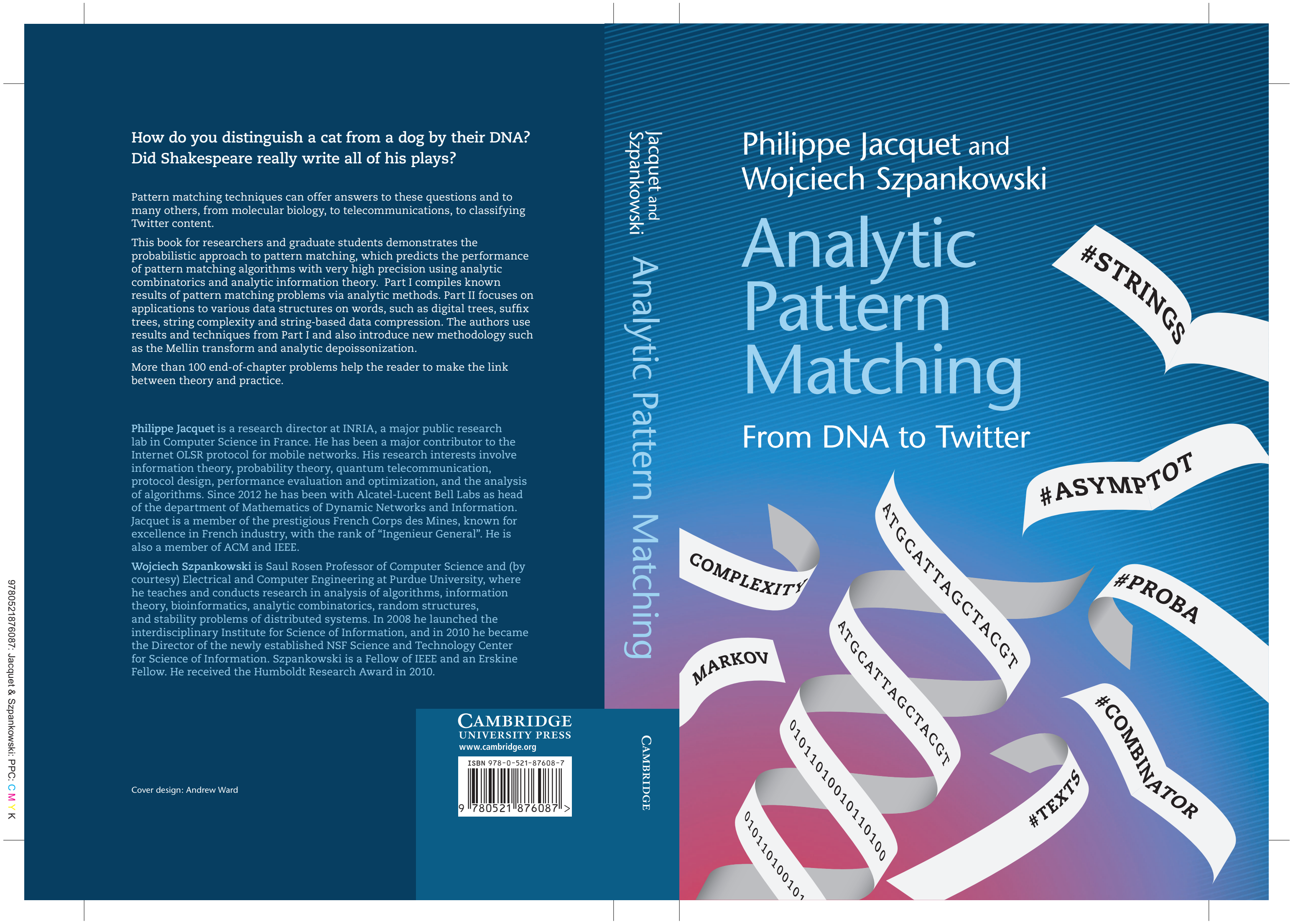
Limited Computational Resources:
In many scenarios, information is limited by available computational resources (e.g., cell phone, living cell).
(Helman & Cover, 1970, “Learning with Limited Memory”.)



Representation-invariant of information:
How to know whether two representations of the same information are information equivalent?



New Book on Pattern Matching



How do you distinguish a cat from a dog by their DNA? Did Shakespeare really write all of his plays?

Pattern matching techniques can offer answers to these questions and to many others, from molecular biology, to telecommunications, to classifying Twitter content.

This book for researchers and graduate students demonstrates the probabilistic approach to pattern matching, which predicts the performance of pattern matching algorithms with very high precision using analytic combinatorics and analytic information theory. Part I compiles known results of pattern matching problems via analytic methods. Part II focuses on applications to various data structures on words, such as digital trees, suffix trees, string complexity and string-based data compression. The authors use results and techniques from Part I and also introduce new methodology such as the Mellin transform and analytic depoissonization.

More than 100 end-of-chapter problems help the reader to make the link between theory and practice.

Philippe Jacquet is a research director at INRIA, a major public research lab in Computer Science in France. He has been a major contributor to the Internet OLSR protocol for mobile networks. His research interests involve information theory, probability theory, quantum telecommunication, protocol design, performance evaluation and optimization, and the analysis of algorithms. Since 2012 he has been with Alcatel-Lucent Bell Labs as head of the department of Mathematics of Dynamic Networks and Information. Jacquet is a member of the prestigious French Corps des Mines, known for excellence in French industry, with the rank of "Ingenieur General". He is also a member of ACM and IEEE.

Wojciech Szpankowski is Saul Rosen Professor of Computer Science and (by courtesy) Electrical and Computer Engineering at Purdue University, where he teaches and conducts research in analysis of algorithms, information theory, bioinformatics, analytic combinatorics, random structures, and stability problems of distributed systems. In 2008 he launched the interdisciplinary Institute for Science of Information, and in 2010 he became the Director of the newly established NSF Science and Technology Center for Science of Information. Szpankowski is a Fellow of IEEE and an Erskine Fellow. He received the Humboldt Research Award in 2010.

Cover design: Andrew Ward

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-0-521-87608-7

9780521876087: Jacquet & Szpankowski: PPC: C M Y K

Jacquet and
Szpankowski
Analytic Pattern Matching
CAMBRIDGE

Philippe Jacquet and
Wojciech Szpankowski

Analytic
Pattern
Matching

From DNA to Twitter

Book Contents

Selected Quotes from Letter-Writers

- “Many of Szpankowski’s research works are jewels of discrete mathematics... Few senior researchers can avail themselves of being at the origin of a new field and few have such a splendid record of original results.” (Flajolet)
- “Prof. Szpankowski distinguishes himself not only by his broad and deep technical contributions, but also by his scientific vision... I believe Prof. Szpankowski brings great distinction to Purdue University, and I cannot think of a more appropriate person to receive this award.” (Kumar)
- “I consider his work on the analysis of Lempel-Ziv compression schemes in information theory his best. Just for this alone, I would not be surprised to see him capture one day the Shannon award... am honored and proud to support Wojtek for a Purdue University Research and Scholarship Distinction Award in Pure or Applied Science or Engineering.” (Devroye)
- “There are few researchers in the field of computer science whose research has had the impact of Wojciech Szpankowski. I am honored to be asked to support his nomination.” (Sedgewick)

Thank You!