Shannon Information Theory and Beyond

W. Szpankowski
Department of Computer Science
Purdue University

November 8, 2015

Center for Science of Information
NSF Science & Technology Center

Arden L. Bement Jr. Lecture, Purdue, 2015

Outline

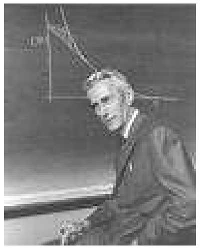
1. Shannon Legacy
   - What is Information?
   - Three Jewels of Shannon
2. Post-Shannon
   - Science of Information
   - Challenges
3. Technical Contribution
   - Constrained Channel Capacity
   - Structural Information and Graph Compression

Shannon Legacy

The Information Revolution started in 1948, with the publication of:

A Mathematical Theory of Communication.

The digital age began.



Claude Shannon:

Shannon $\boxed{\text{information}}$ quantifies the extent to which a recipient of data can reduce its statistical uncertainty.

"These semantic aspects of communication are irrelevant . . ."

$\boxed{\text{Fundamental Limits for \textbf{Compression} and \textbf{Data Transmission}.}}$

Applications Enabler/Driver:

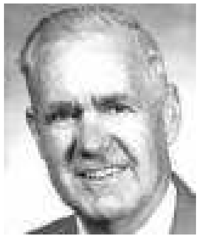
CD, iPod, DVD, video games, computer communication, Internet, Facebook, Google, . . .

Design Driver:

universal data compression, data encoding, voiceband modems, CDMA, multiantenna, discrete denosing, space-time codes, cryptography, . . .

What is Information?

C. F. Von Weizsäcker:



"Information is only that which produces information" (relativity).
"Information is only that which is understood" (rationality)
"Information has no absolute meaning".

R. Feynman:

". . . Information is as much a property of your own knowledge as anything in the message.
. . . Information is not simply a physical property of a message: it is a property of the message and your knowledge about it."

J. Wheeler:
"It from Bit". (Information is physical.)

A. Zeilinger:
. . . reality and information are two sides of the same coin, that is, they are in a deep sense indistinguishable.

What is Information?

Information has the flavor of:

relativity (depends on the activity undertaken),

rationality (depends on the recipient's knowledge),

timeliness (temporal structure),

space (spatial structure).

Informally Speaking: A piece of data carries **information** if it can impact a **recipient's ability** to achieve the **objective** of some **activity** within a given **context**.

What is Information?

Information has the flavor of:

relativity (depends on the activity undertaken),

rationality (depends on the recipient's knowledge),

timeliness (temporal structure),

space (spatial structure).

Informally Speaking: A piece of data carries **information** if it can impact a **recipient's ability** to achieve the **objective** of some **activity** within a given **context**.

Engineering ViewPoint:

> Information is a measure of distinguishability.

What is Information?

Information has the flavor of:

relativity (depends on the activity undertaken),

rationality (depends on the recipient's knowledge),

timeliness (temporal structure),

space (spatial structure).

Informally Speaking: A piece of data carries **information** if it can impact a **recipient's ability** to achieve the **objective** of some **activity** within a given **context**.

Engineering ViewPoint:

> Information is a measure of distinguishibility.

Example: Boltzmann's Question:
Among many possible gas molecules (distinguishable) distributions, which one is the most likely to occur?

Shannon Information . . .

In our setting, Shannon defined:

objective: statistical ignorance of the recipient;
statistical uncertainty of the recipient.

cost: # binary decisions to describe E;
= $-\log P(E)$; $P(E)$ being the probability of E.

Context: the semantics of data is irrelevant . . .

Self-information for E_i: $\mathrm{info}(E_i) = -\log P(E_i)$.
Average information: $H(P) = -\sum_i P(E_i) \log P(E_i)$
Entropy of $X = \{E_1, \ldots\}$: $H(X) = -\sum_i P(E_i) \log P(E_i)$
Mutual Information: $I(X;Y) = H(Y) - H(Y|X)$, (faulty channel).

Shannon's statistical information tells us how much a recipient of data can reduce their statistical uncertainty by observing data.

Shannon's information is not absolute information since $P(E_i)$ (prior knowledge) is a subjective property of the recipient.

Shortest Description, Complexity

Example: X can take eight values with probabilities:

$$P = (p_1, \dots, p_8) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{16}, \tfrac{1}{64}, \tfrac{1}{64}, \tfrac{1}{64}, \tfrac{1}{64}).$$

Assign to them the following code:

$$0, 10, 110, 1110, 111100, 111101, 111110, 111111,$$

The length of this code $L(X)$ (shortest description):

$$L(X) = \sum_{i=1}^{8} p_i l_i = 2 \text{ bits}.$$

and entropy X

$$H(X) = 2 \text{ bits}.$$

In general, if X is a (random) sequence with entropy $H(X)$ and average code length $L(X)$, then

$$H(X) \leq L(X) \leq H(X) + 1.$$

Complexity vs Description vs Entropy
The more complex X is, the longer its description is, and the bigger the entropy is.

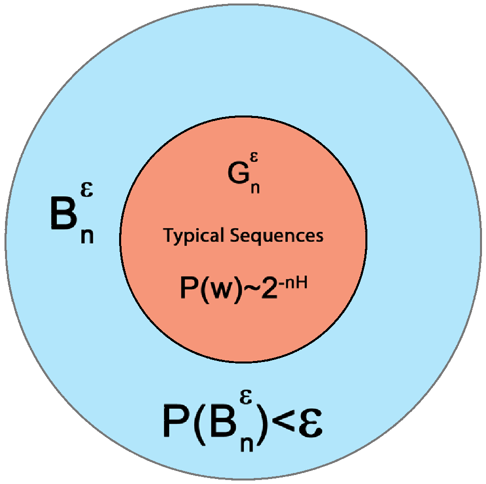
Three Theorems of Shannon

How many bits (minimum) you need to describe a file?

Theorem 1 & 3. (**Shannon 1948; Lossless & Lossy Data Compression**)

compression bit rate $\geq$ source entropy $H(X)$

for distortion level D:
lossy bit rate $\geq$ rate distortion function $R(D)$



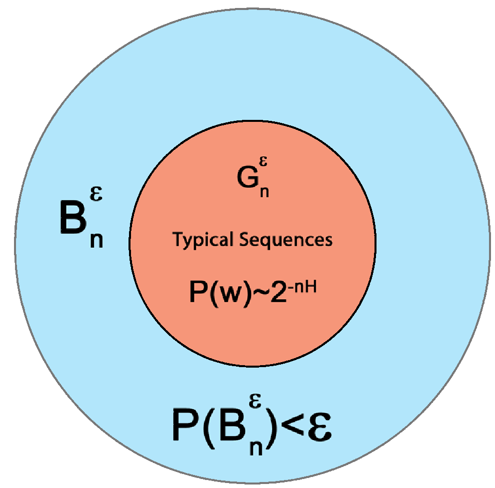
Three Theorems of Shannon

How many bits (minimum) you need to describe a file?

Theorem 1 & 3. (**Shannon 1948; Lossless & Lossy Data Compression**)

compression bit rate $\geq$ source entropy $H(X)$

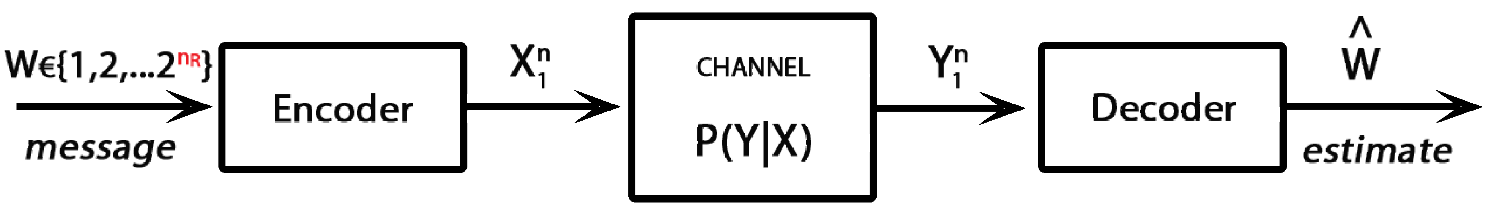
for distortion level D:
lossy bit rate $\geq$ rate distortion function $R(D)$



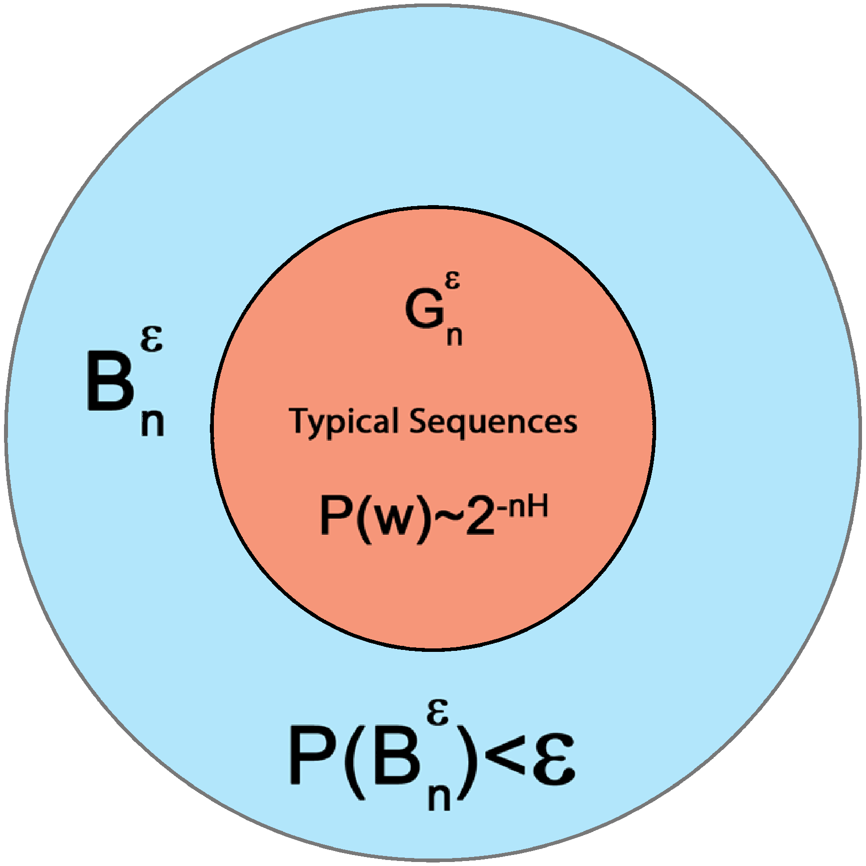
How many bits we can communicate reliably over a noisy channel?

Theorem 2. (**Shannon 1948; Channel Coding**)
In Shannon's words:

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.



Typical Sequences



Shannon-McMilan-Breiman:

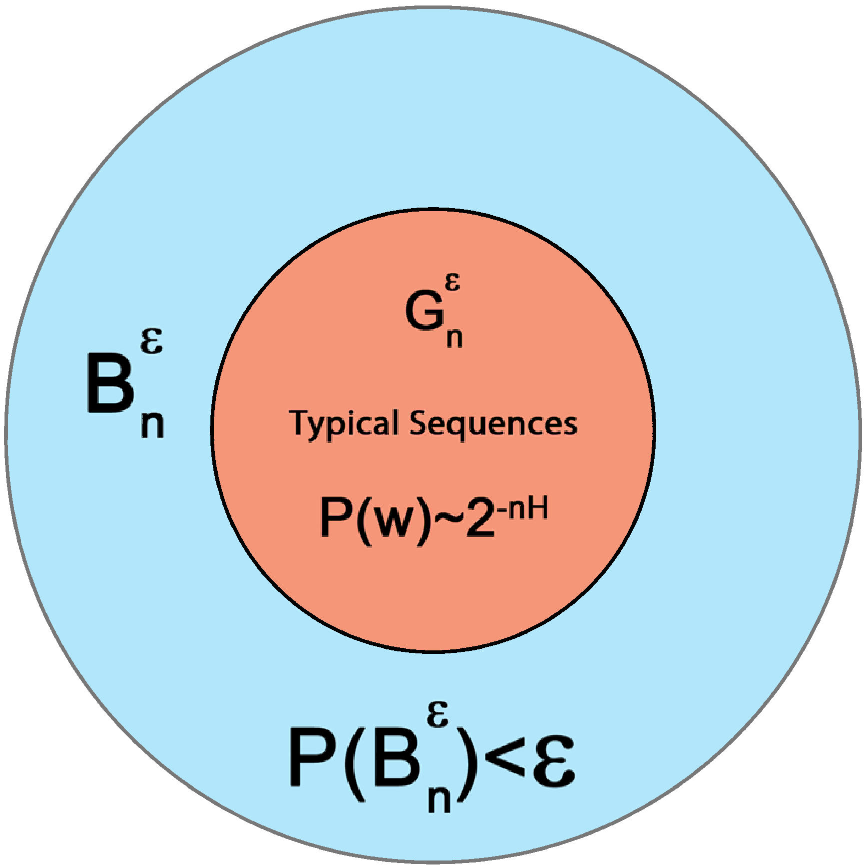
$$-\frac{1}{n}\log P(X_1^n) \to H(X) = -\mathbf{E}[\log P(X)]$$

$H(X)$ is the entropy rate.

Code Length :

$$\lceil -\log P(X_1^n) \rceil \sim nH(X).$$

Typical Sequences



Shannon-McMilan-Breiman:

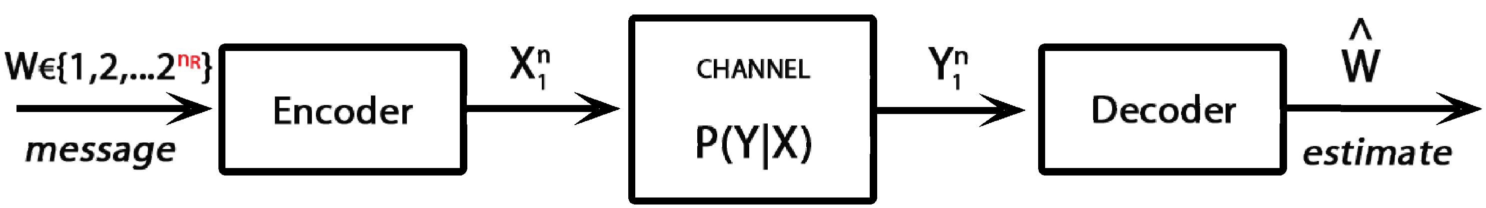
$$-\frac{1}{n}\log P(X_1^n) \to H(X) = -\mathbf{E}[\log P(X)]$$

$H(X)$ is the entropy rate.

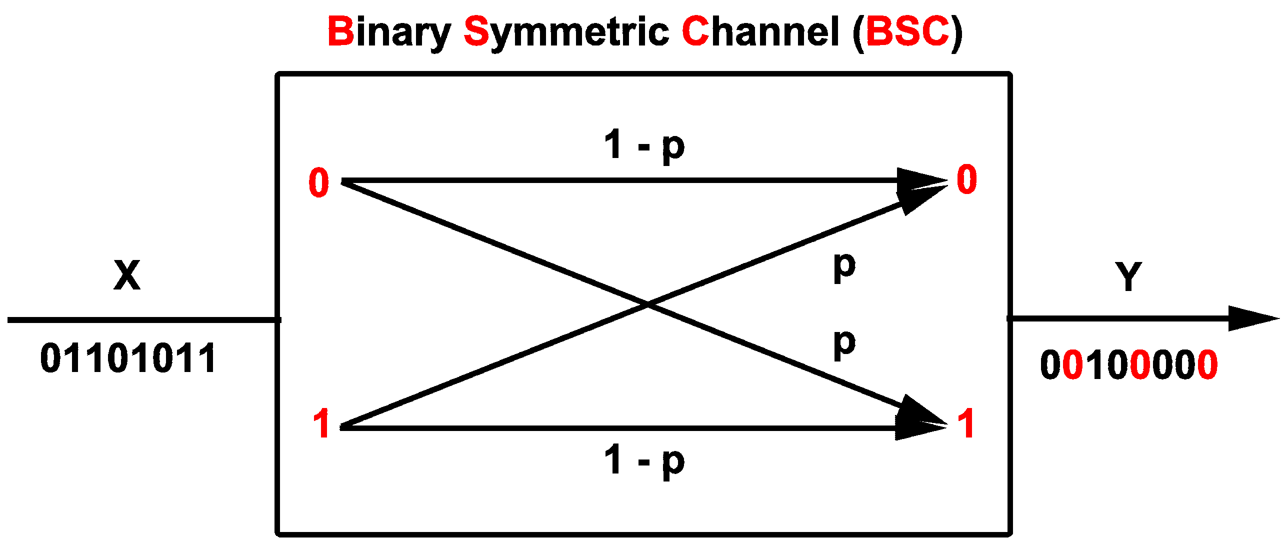
Code Length :

$$\lceil -\log P(X_1^n) \rceil \sim nH(X).$$

Decoding Rule: Declare that sequence sent X is the one that is jointly typical with the received sequence Y provided there is unique X satisfying this property!



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

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

Outline Update

1. Shannon Legacy
2. Post-Shannon
   - Science of Information
   - Challenges
3. Technical Contribution

Shannon Information vs Science of Information

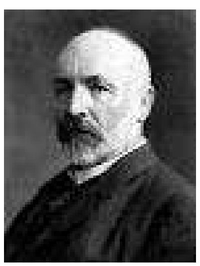
Claude Shannon laid the foundation of information theory, demonstrating that problems of data transmission and compression (i.e., reliably **reproducing data**) can be precisely modeled formulated, and analyzed.

SCIENCE OF INFORMATION builds on Shannon's principles to address key challenges in understanding information that nowadays is not only communicated but also acquired, curated, organized, aggregated, managed, processed, suitably abstracted and represented, analyzed, inferred, valued, secured, and used in various scientific, engineering, and socio-economic processes

Shannon Information vs Science of Information

Claude Shannon laid the foundation of information theory, demonstrating that problems of data transmission and compression (i.e., reliably **reproducing data**) can be precisely modeled formulated, and analyzed.

SCIENCE OF INFORMATION builds on Shannon's principles to address key challenges in understanding information that nowadays is not only communicated but also acquired, curated, organized, aggregated, managed, processed, suitably abstracted and represented, analyzed, inferred, valued, secured, and used in various scientific, engineering, and socio-economic processes



Gergor Cantor (1845-1918):
"*In re mathematica ars proponendi questionem pluris facienda est quam solvendi*"
(In mathematics the art of proposing a question must be held of higher value than solving it.)

Post-Shannon Challenges

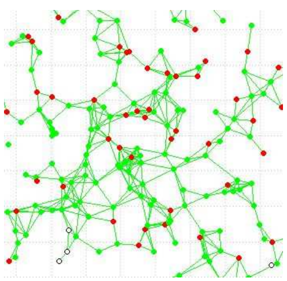
Classical Information Theory needs a recharge to meet new challenges of nowadays applications in biology, modern communication, knowledge extraction, economics and physics, ...

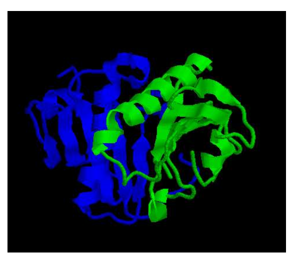
We need to extend Shannon information theory to include new aspects of information such as:

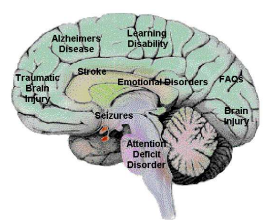
structure, time, space, and semantics ,

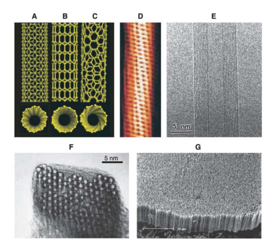
and others such as:

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









Outstanding Challenges in Science of Information

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

$$\text{data} \rightarrow \text{information} \rightarrow \text{knowledge}$$

Outstanding Challenges in Science of Information

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

$$\text{data} \rightarrow \text{information} \rightarrow \text{knowledge}$$

1. **Easy Questions**: How much unique data?
   Increasingly data is not in the form of text – social networks, tweets, scientific data (interactions, geometries, time series), economic transactions, etc.

2. **Harder Questions**: How do we quantify this data, how do we extract information from these datasets?

3. **Really Hard Questions**: Information has cause and consequence – How do we reach beyond information? How do we act on this information?

Outline Update

1. Shannon Legacy
2. Post-Shannon Challenges
3. Technical Contribution
   - Constrained Channel Capacity
   - Structural Information and Graph Compression

Channel with Constrained Input

In many real applications (such as digital recording and biology), input sequence must satisfy some constrains such as (d, k) sequences:

No sequence contains a run of zeros shorter than d or longer than k.

Digital Recording such as CD, DVD, and Blu-ray:
An unconstrained sequence of $1'$s and $0'$s is not acceptable in practice. since a long run of $0'$s results in loss of synchronization. Therefore, constrained (d, k) sequences are used to improve the performance.

Neuronal Spike
Current technology allows for the simultaneous recording of the spike trains from one hundred different neurons in the brain of a live animal. But refractoriness requires that a neuron cannot fire two spikes in too short a time, thus constrained (d, k) sequences arise.

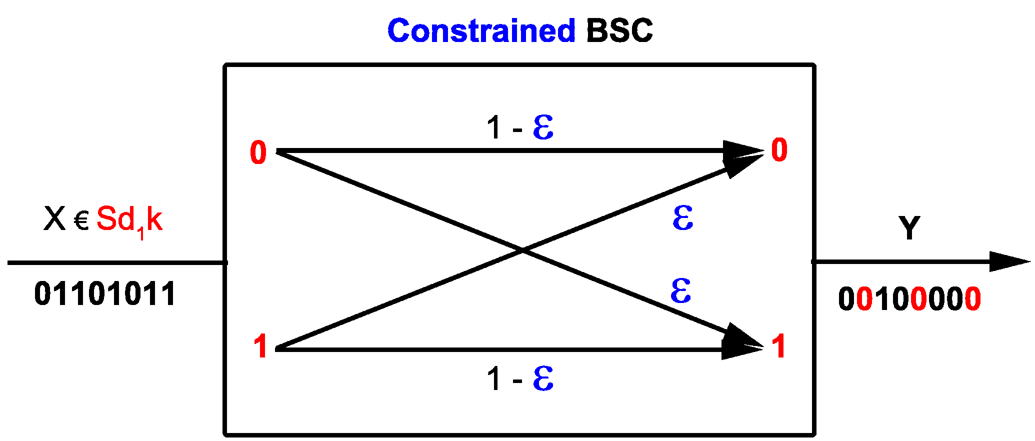
Channel with Constrained Input

In many real applications (such as digital recording and biology), input sequence must satisfy some constrains such as (d, k) sequences:

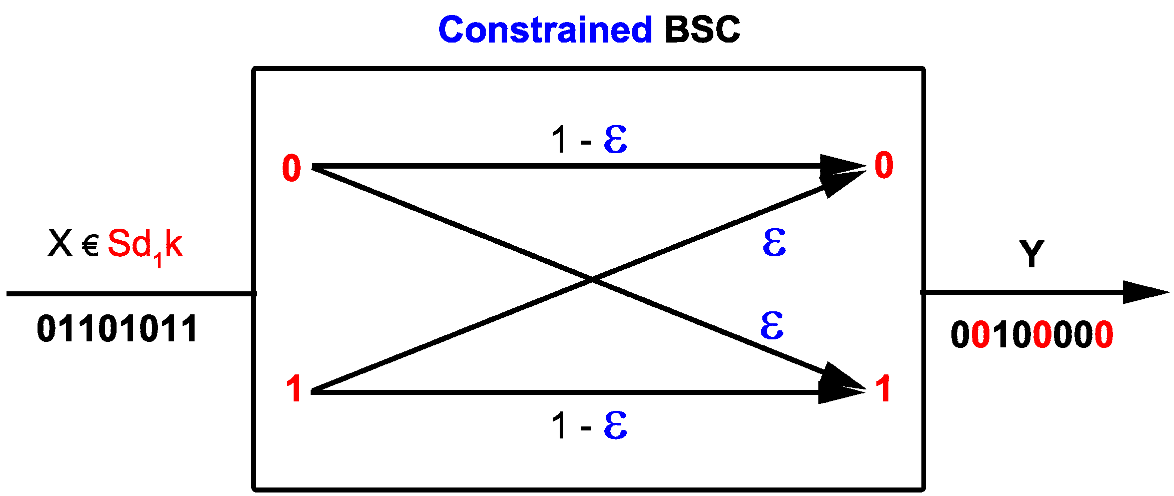
$$\boxed{\text{No sequence contains a run of zeros shorter than } d \text{ or longer than } k.}$$

Digital Recording such as CD, DVD, and Blu-ray:
An unconstrained sequence of 1's and 0's is not acceptable in practice. since a long run of 0's results in loss of synchronization. Therefore, constrained (d, k) sequences are used to improve the performance.

Neuronal Spike
Current technology allows for the simultaneous recording of the spike trains from one hundred different neurons in the brain of a live animal. But refractoriness requires that a neuron cannot fire two spikes in too short a time, thus constrained (d, k) sequences arise.



Noisy Constrained Channel

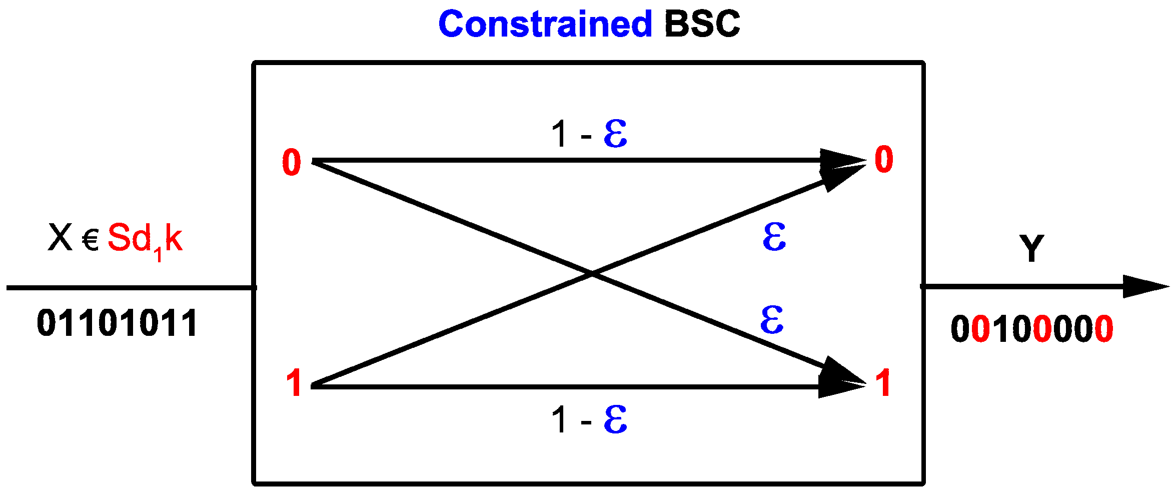


Let $\mathcal{S}$ denote the set of binary constrained sequences of length n. Here:

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

Sequence $X \in \mathcal{S}_{(d,k)}$ is a **MARKOV PROCESS** of order k.

Noisy Constrained Channel



Let $\mathcal{S}$ denote the set of binary constrained sequences of length n. Here:

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

Sequence $X \in \mathcal{S}_{(d,k)}$ is a **MARKOV PROCESS** of order k.

Capacity:

$C(\mathcal{S}, \varepsilon)$ – 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 was an **open problem**.[1]

[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 . . .*"

Entropy of Hidden Markov Process

Hidden Markov Process: Since

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

Process Y is a Hidden Markov Process (HMP) since it is a noisy version of the Markov Process X.

Entropy of HMP $H(Y)$ was first investigated by Blackwell in 1956.
We proved that $H(Y)$ is equal to the so called top Lyapunov exponent which is hard to compute.

Entropy of Hidden Markov Process

Hidden Markov Process: Since

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

Process Y is a Hidden Markov Process (HMP) since it is a noisy version of the Markov Process X.

Entropy of HMP $H(Y)$ was first investigated by Blackwell in 1956.
We proved that $H(Y)$ is equal to the so called top Lyapunov exponent which is hard to compute.

We now assume that $P(\text{error}) = \varepsilon \to 0$ is small!

Theorem 1 (Jacquet, Seroussi, and Szpankowski, 2008). *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 $f_1(P)$.

Entropy of Hidden Markov Process

Hidden Markov Process: Since

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

Process Y is a Hidden Markov Process (HMP) since it is a noisy version of the Markov Process X.

Entropy of HMP $H(Y)$ was first investigated by Blackwell in 1956.
We proved that $H(Y)$ is equal to the so called top Lyapunov exponent which is hard to compute.

We now assume that $P(\mathrm{error}) = \varepsilon \to 0$ is small!

Theorem 1 (Jacquet, Seroussi, and Szpankowski, 2008). *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 $f_1(P)$.

Example 1: Consider $\mathbf{P} = \begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix}$ which represents $(0,1) \equiv (1,\infty)$ constraint sequence. Then

$$H(Y) = H(P) - \frac{p(2-p)}{1+p}\varepsilon \log \varepsilon + O(\varepsilon).$$

Capacity of the Noisy Constrained Channel

Theorem 2 (Jacquet & Szpankowski, 2010)**.** *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 (*$\varepsilon=0$*)*

Capacity of the Noisy Constrained Channel

Theorem 2 (Jacquet & Szpankowski, 2010). *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 ($\varepsilon = 0$)

Example 2. Consider the $(1,\infty) \equiv (0,1)$ constraint (at most one 0 between any two 1s) with transition matrix as in Example 1. Then

$$f_0(P_X) = \frac{p(p-2)}{p-1},$$

The noisy constrained capacity is obtained for:
$p = 1/\varphi^2$, where $\varphi = (1+\sqrt{5})/2$, (the golden ratio). Then

$$\begin{aligned} C(\mathcal{S},\varepsilon) &= C(\mathcal{S}) + +(1 - 1/\sqrt{5})\varepsilon \log(\varepsilon) + O(\varepsilon) \\ &= \log \varphi + (1 - 1/\sqrt{5})\varepsilon \log(\varepsilon) + O(\varepsilon) \end{aligned}$$

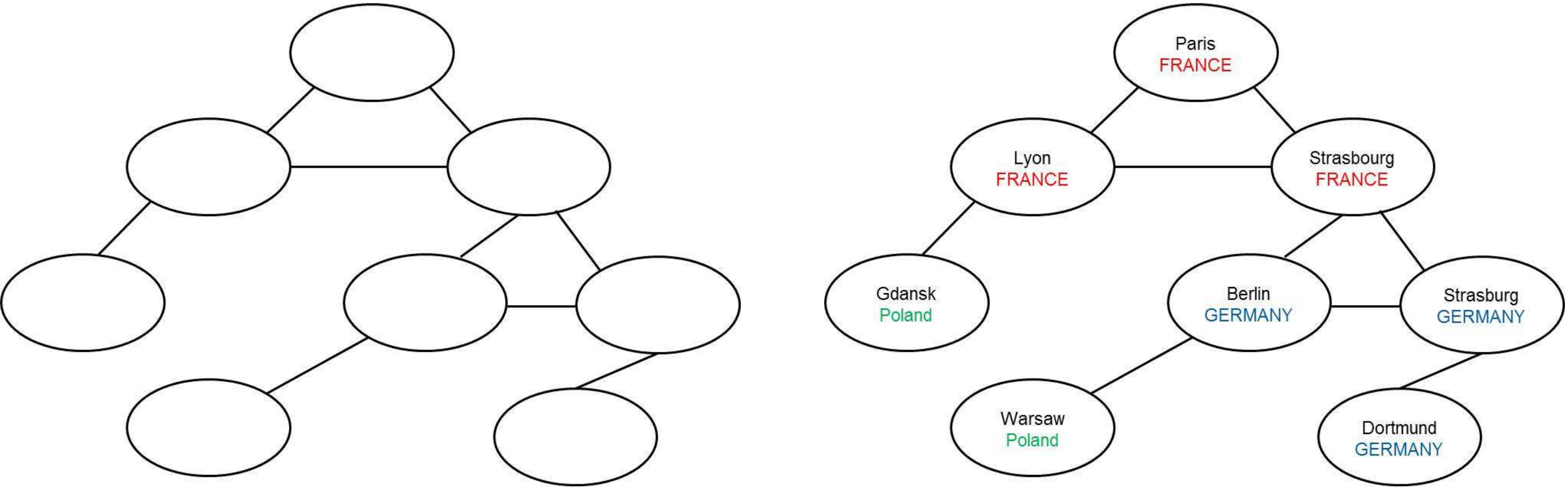
for $\varepsilon \to 0$.

Outline Update

1. Shannon Legacy
2. Post-Shannon Challenges
3. Technical Contribution
   - Constrained Channel Capacity
   - Structural Information and Graph Compression[2]"

[2] F. Brooks. jr "We have no theory however that gives us a metric for the information embodied in structure . . .

Graphs with Locally Correlated Labels



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?

The Real Stuff ...

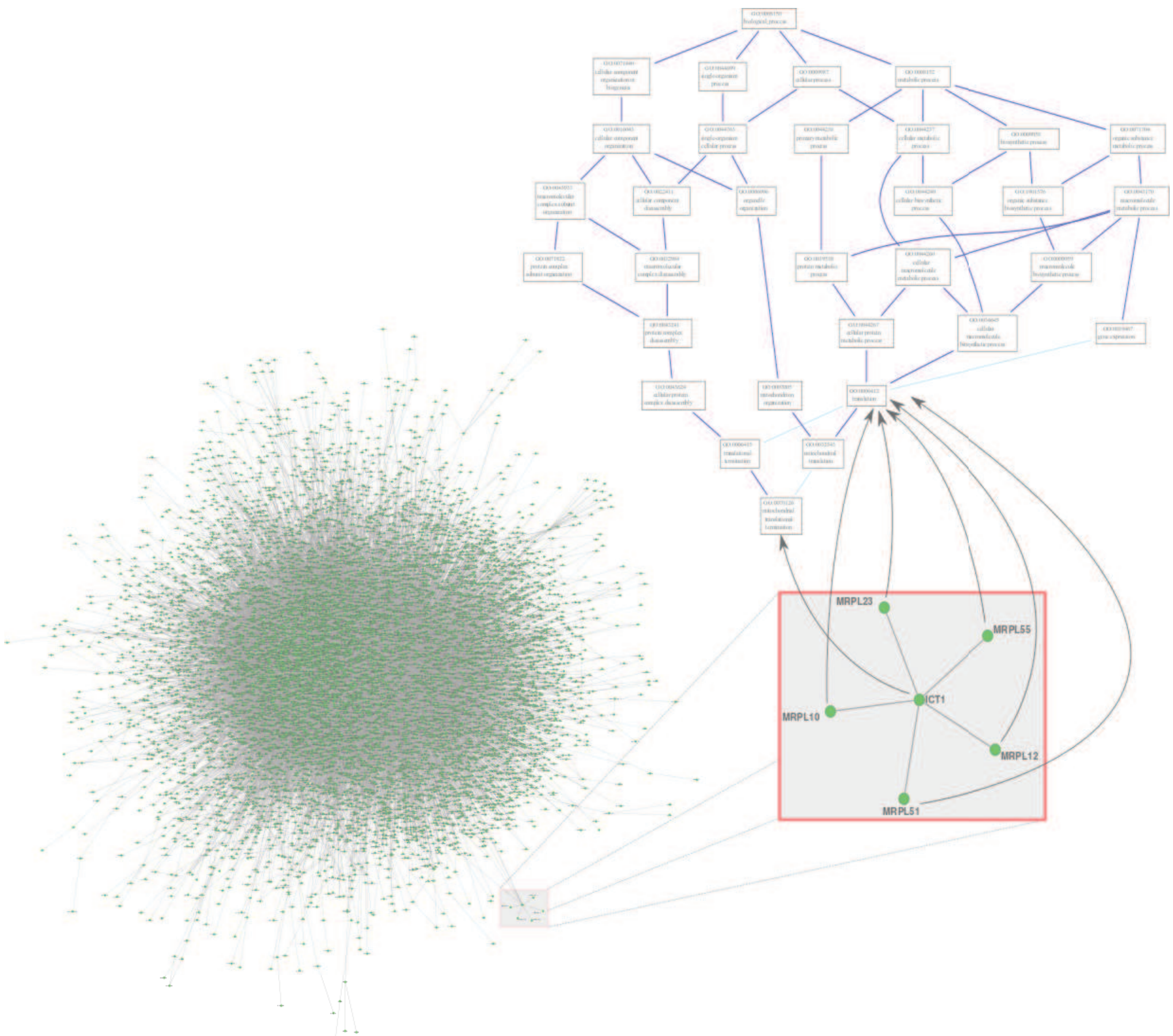
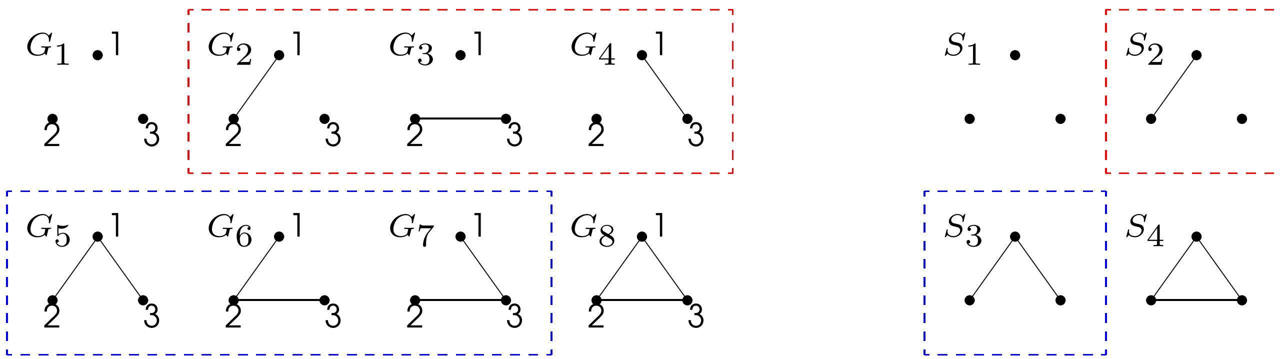


Figure 1: Protein-Protein Interaction Network with BioGRID database

Graph and Structural Entropies

Information Content of Unlabeled Graphs:

A structure model $\mathcal{S}$ of a graph $\mathcal{G}$ is defined for an unlabeled version.
Some labeled graphs have the same structure.



Graph Entropy vs **Structural Entropy**:

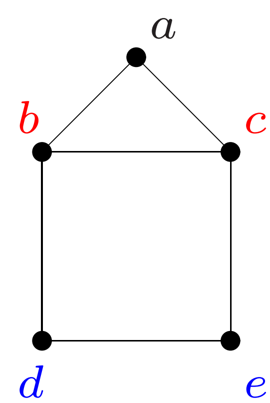
The probability of a structure S is: $P(S) = N(S) \cdot P(G)$
where $N(S)$ is the number of different labeled graphs having the same structure.

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

$$H_{\mathcal{S}} = \mathbf{E}[-\log P(S)] = -\sum_{S \in \mathcal{S}} P(S) \log P(S) \quad \text{structural entropy}$$

Relationship between $H_{\mathcal{G}}$ and $H_{\mathcal{S}}$

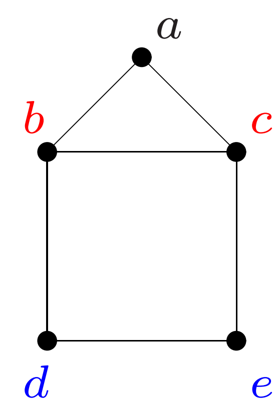
Graph Automorphism: For a graph G its automorphism $\mathrm{Aut}(G)$ is adjacency preserving permutation of vertices of G.



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

automorphism group of S.

Relationship between $H_{\mathcal{G}}$ and $H_{\mathcal{S}}$



Graph Automorphism: For a graph G its automorphism $\mathrm{Aut}(G)$ is adjacency preserving permutation of vertices of G.

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

automorphism group of S.

Erdös-Rényi Graph Model: Such a graph $\mathcal{G}(n, p)$ with n vertices edges are chosen independently with probability p. That is,

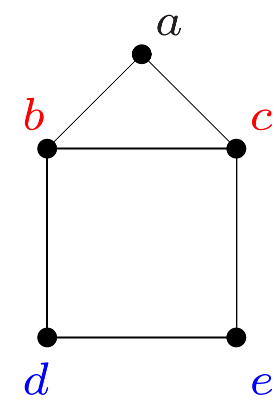
$$P(G) = p^k((1-p)^{\binom{n}{2}-k}$$

and Kim, Sudakov, Vu (2006) prove that for such graphs

$$P(\mathrm{Aut}(G) = 1) = 1 - o(1).$$

Relationship between $H_{\mathcal{G}}$ and $H_{\mathcal{S}}$

Graph Automorphism: For a graph G its automorphism $\mathrm{Aut}(G)$ is adjacency preserving permutation of vertices of G.



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

automorphism group of S.

Erdös-Rényi Graph Model: Such a graph $\mathcal{G}(n,p)$ with n vertices edges are chosen independently with probability p. That is,

$$P(G) = p^k((1-p)^{\binom{n}{2}-k}$$

and Kim, Sudakov, Vu (2006) prove that for such graphs

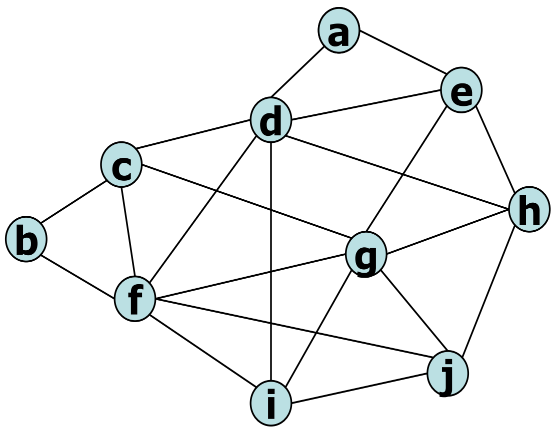
$$P(\mathrm{Aut}(G) = 1) = 1 - o(1).$$

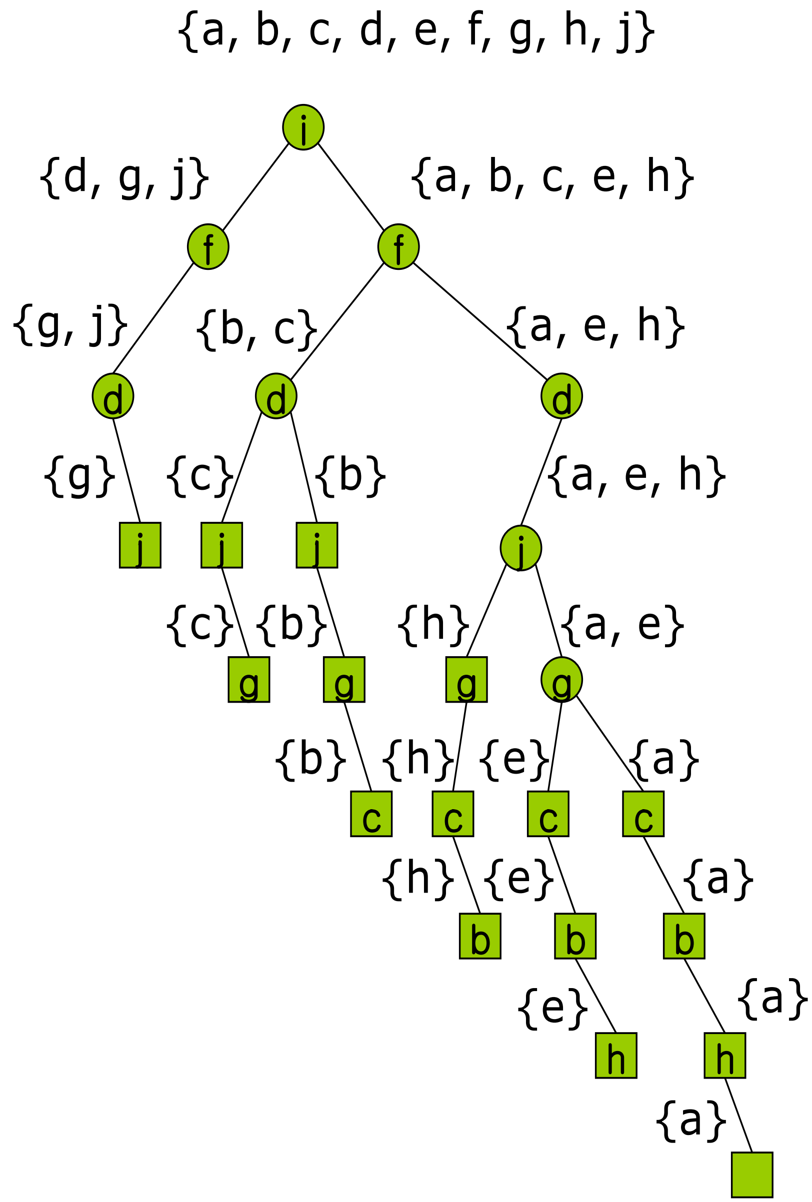
Theorem 3 (Choi and Szpankowski, 2012). *For large n and all p satisfying $\frac{\ln n}{n} \ll p$ and $1 - p \gg \frac{\ln n}{n}$ (i.e., the graph is connected w.h.p.),*

$$H_{\mathcal{S}} = \binom{n}{2} h(p) - \log n! + O\left(\frac{\log n}{n^a}\right) = \binom{n}{2} h(p) - n \log n + n \log e + O(\log n), \quad a > 1$$

where $h(p) = -p \log p - (1-p) \log(1-p)$ is the entropy rate.

Structural Zip (SZIP) Algorithm





B1 = 0100110100001110101

B2 = 100101100000101

Asymptotic Optimality of SZIP for Erdös-Rényi Graphs

Theorem 4 (Y. Choi and W. Szpankowski, 2012). *Let* $L(S) = |\tilde{B}_1| + |\tilde{B}_2|$ *be the code length.*

(i) *For large* n*,*

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

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

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

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

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

Asymptotic Optimality of SZIP for Erdös-Rényi Graphs

Theorem 4 (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* $\varepsilon > 0$*,*

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

(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

Analytic Information Theory

- In the **1997 Shannon Lecture** Jacob Ziv presented compelling arguments for "backing off" from first-order asymptotics in order to predict the behavior of real systems with finite length description.

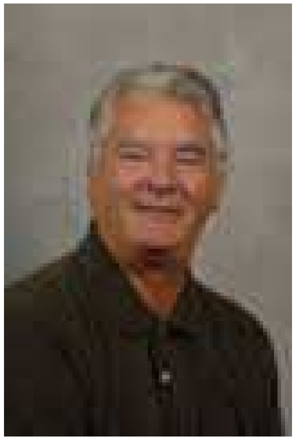
- Following **Hadamard's precept**[3], we study information theory problems using techniques of complex analysis[4] such as generating functions, combinatorial calculus, Rice's formula, Mellin transform, Fourier series, sequences distributed modulo 1, saddle point methods, analytic poissonization and depoissonization, and singularity analysis.

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

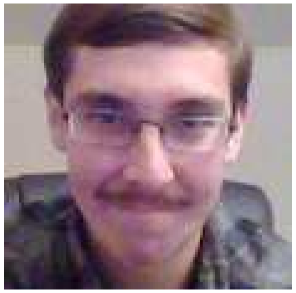
[3] The shortest path between two truths on the real line passes through the complex plane.

[4] Andrew Odlyzko argued that: "*Analytic methods are extremely powerful and when they apply, they often yield estimates of unparalleled precision.*"

Acknowledgments

. . . and my current and former students

That's It