From Local Information to Global Inference

Wojciech Szpankowski
Department of Computer Science
Center for Science of Information (CSoI)
Purdue University
W. Lafayette, IN 47907

January 23, 2020

CSoI Workshop, Hawaii, 2020

Outline

1. Workshop Objectives

2. NSF Center for Science of Information (CSoI)

3. Examples
   - Analytic Combinatorics: (Redundancy Problem of Source Coding)
   - Graphs: (Shortest Common Superstring)
   - Structure: (Graph Compression)

Algorithms:	are at the heart of virtually all computing technologies;
Combinatorics:	provides indispensable tools for finding patterns and structures;
Information:	measure of distinguishibility.

Objectives of the Workshop

Goals: *Principled investigation of the coupling between what may be termed the "local information" and more traditional questions about its global feasibility/complexity/inference. Our quest is for radical new ways of reasoning about the local-to-global nature of information and computation*

Examples. **1**. In social networks there is often substantial "small scale" structure (e.g., clusters in so-called ego networks), but the global properties of the adjacency matrix or Laplacian matrix are more consistent with the hypothesis of unstructured noise.

2. In neural networks training overly-parameterized models against large quantities of data leads to localized structures in weight matrices that combine to achieve high quality (but brittle) models.

3. In data science consistency of model classes is very important. To enable handling rich probabilistic model classes we must study a data-driven consistency framework. It shifts focus from the global complexity of the class to a form of local complexity that capture the local variation of properties within the model classes by means of topological formulations.

4. In problems where one finds an analytic representation through a complex function, a singular local point determines global asymptotics.

Outline

1. Workshop Objectives
2. **NSF Center for Science of Information** (CSoI)
3. Examples

NSF Center for Science of Information

In 2010 National Science Foundation established

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: Berkeley, MIT, Hawaii, Princeton, Stanford, Texas A&M, UIUC, UCSD and Bryn Mawr & Howard U.

Specific Center's Goals:

- define core theoretical principles governing transfer of information.
- develop meters and methods for information.
- apply to problems in physical and social sciences, and engineering.
- offer a venue for multi-disciplinary long-term collaborations.
- transfer advances in research to education and industry.

Post-Shannon Challenges

1. **Back off from infinity** (Ziv'97): Extend Shannon findings to finite size data structures (i.e., sequences, graphs), that is, develop information theory of various data structures beyond first-order asymptotics.

> **Claim**: Many interesting information-theoretic phenomena appear in the second-order terms.

2. **Science of Information**: Information Theory needs to meet new challenges of current applications in

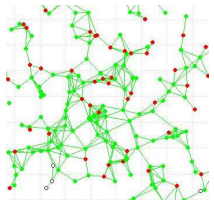
biology, communication, knowledge extraction, economics, . . .

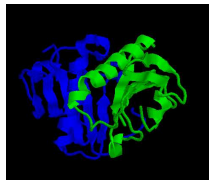
to understand new aspects of information in:

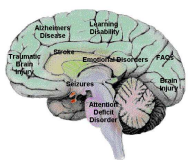
structure, time, space, and semantics,

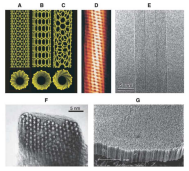
and

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









Value Added and Legacy

Value Added and Legacy:

1. Legacy of new collaboration between different disciplines (biology, chemistry, computer science, information theory, and statistics)
2. Legacy of new collaboration within the same discipline (e.g., genomic compression, coded string reconstruction)
3. Legacy of educating new crops of researchers (Courtade, Grover, Kostina, Oshman, Polyanskiy)
4. Legacy of new research directions (information theory in life sciences, information theory in data science, community detection, information theory in discrete geometry)
5. Legacy of formulating new foundations: (security, structure, temporal, dynamic networks, privacy)

Outline

1. Workshop Objectives

2. NSF Center for Science of Information (CSoI)

3. Examples
   - Analytic Combinatorics: (**Redundancy Problem of Source Coding**)
   - Graphs: (Shortest Common Superstring)
   - Structure: (Graph Compression)

Source Coding and Redundancy

Source coding aims at finding codes $C : \mathcal{A}^* \to \{0,1\}^*$ of the shortest length $L(C,x)$, either on *average* or for *individual sequences*.

Known Source P: The pointwise and maximal redundancy are:

$$\begin{aligned} R_n(C_n, P; x_1^n) &= L(C_n, x_1^n) + \log P(x_1^n) \\ R^*(C_n, P) &= \max_{x_1^n} [L(C_n, x_1^n) + \log P(x_1^n)] \end{aligned}$$

where $P(x_1^n)$ is the probability of $x_1^n = x_1 \cdots x_n$.

Unknown Source P: Following Davisson, the maximal minimax redundancy $R_n^*(\mathcal{S})$ for a family of sources $\mathcal{S}$ is:

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

Shtarkov's Bound:

$$d_n(\mathcal{S}) := \log \sum_{x_1^n \in \mathcal{A}^n} \sup_{P \in \mathcal{S}} P(x_1^n) \leq R_n^*(\mathcal{S}) \leq \log \underbrace{\sum_{x_1^n \in \mathcal{A}^n} \sup_{P \in \mathcal{S}} P(x_1^n)}_{D_n(\mathcal{S})} + 1$$

Maximal Minimax for Memoryless Sources

For a **memoryless source** over the alphabet $\mathcal{A} = \{1, 2, \ldots, m\}$ we have

$$P(x_1^n) = {p_1}^{k_1} \cdots {p_m}^{k_m}, \qquad k_1 + \cdots + k_m = n.$$

Then

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

since the (unnormalized) **likelihood distribution** is

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

Generating Function for $D_n(\mathcal{M}_0)$

We write

$$D_n(\mathcal{M}_0) = \sum_{k_1+\cdots+k_m=n} \binom{n}{k_1,\ldots,k_m} \left(\frac{k_1}{n}\right)^{k_1} \cdots \left(\frac{k_m}{n}\right)^{km} = \frac{n!}{n^n} \sum_{k_1+\cdots+k_m=n} \frac{k_1^{k_1}}{k_1!} \cdots \frac{k_m^{km}}{k_m!}$$

Let us introduce a tree-generating function

$$B(z) = \sum_{k=0}^{\infty} \frac{k^k}{k!} z^k = \frac{1}{1-T(z)}, \qquad T(z) = \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} z^k$$

where $T(z) = ze^{T(z)}$ $(= -W(-z)$, **Lambert's** W-function) that enumerates all rooted labeled trees. Let now

$$D_m(z) = \sum_{n=0}^{\infty} z^n \frac{n^n}{n!} D_n(\mathcal{M}_0).$$

Then by the convolution formula

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

Asymptotics for FINITE m

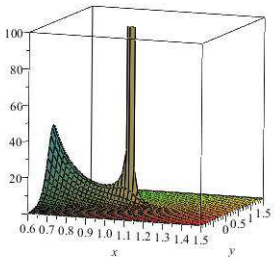
The function $B(z)$ has an algebraic singularity at $z = e^{-1}$, and

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

By Cauchy's coefficient formula

$$D_n(\mathcal{M}_0) = \frac{n!}{n^n}[z^n][B(z)]^m = \sqrt{2\pi n}(1 + O(1/n))\frac{1}{2\pi i}\oint \frac{\beta(z)^m}{z^{n+1}}dz.$$

For finite m, the singularity analysis of Flajolet and Odlyzko implies



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

that finally yields (cf. Clarke & Barron, 1990, W.S., 1998)

$$\begin{aligned} R_n^*(\mathcal{M}_0) &= \frac{m-1}{2}\log\left(\frac{n}{2}\right) + \log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) + \frac{\Gamma(\frac{m}{2})m}{3\Gamma(\frac{m}{2}-\frac{1}{2})}\cdot\frac{\sqrt{2}}{\sqrt{n}} \\ &+ \left(\frac{3 + m(m-2)(2m+1)}{36} - \frac{\Gamma^2(\frac{m}{2})m^2}{9\Gamma^2(\frac{m}{2}-\frac{1}{2})}\right)\cdot\frac{1}{n} + \cdots \end{aligned}$$

Redundancy for LARGE m

Now assume that m is unbounded and may vary with n. Then

$$D_{n,m}(\mathcal{M}_0) = \sqrt{2\pi n}\frac{1}{2\pi i}\oint \frac{\beta(z)^m}{z^{n+1}}dz = \sqrt{2\pi n}\frac{1}{2\pi i}\oint e^{g(z)}dz$$

where $g(z) = m \ln \beta(z) - (n+1)\ln z$.

The saddle point z_0 is a solution of $g'(z_0) = 0$, that is,

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

Under mild conditions satisfied by our $g(z)$ (e.g., z_0 is real and unique), the saddle point method leads to:

$$D_{n,m}(\mathcal{M}_0) = \frac{e^{g(z_0)}}{\sqrt{2\pi|g''(z_0)|}} \times \left(1 + O\left(\frac{g'''(z_0)}{(g''(z_0))^\rho}\right)\right),$$

for some $\rho < 3/2$.

Saddle Point

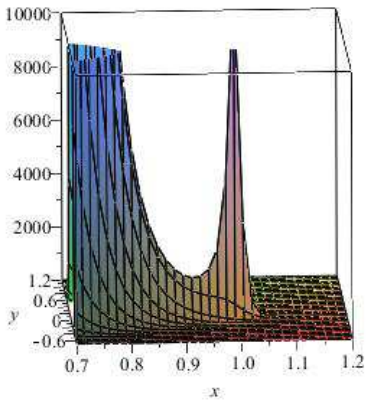
The saddle point z_0 satisfies

$$z_0 \frac{\beta'(z_0)}{\beta(z_0)} = \frac{n+1}{m}.$$

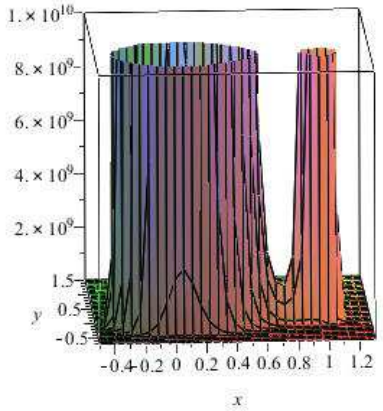
After some algebra we obtain $z_0 = (1-\gamma_{n,m})e^{\gamma_{n,m}}$ where

$$\gamma_{n,m} = \frac{m}{2(n+1)} \left(\sqrt{1 + \frac{4(n+1)}{m}} - 1 \right)$$

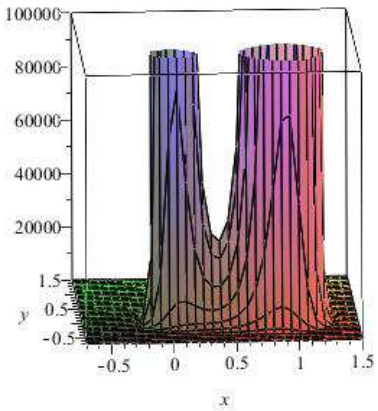
Notice that $0 < z_0 < 1$. More precisely:
(i) $z_0 \to 1$ when $m = o(n)$;
(ii) $0 < z_0 < 1$ when $m = \Theta(n)$;
(iii) $z_0 \to 0$ when $n = o(m)$.



$m = o(n)$



$m = n$



$n = o(m)$

Main Results for LARGE m

Theorem 1 (Orlitsky and Santhanam, 2004, and W.S. and Weinberger, 2010)**.**
(i) *For* $m = o(n)$

$$R^*_{n,m}(\mathcal{M}_0) = \frac{m-1}{2}\log\frac{n}{m} + \frac{m}{2}\log e + \frac{m\log e}{3}\sqrt{\frac{m}{n}}\frac{1}{2} - O\left(\sqrt{\frac{m}{n}}\right)$$

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

$$R^*_{n,m}(\mathcal{M}_0) \quad = \quad n\log B_\alpha + \ell(n)\log C_\alpha - \log\sqrt{A_\alpha} + O(\ell(n)^2/n)$$

where $C_\alpha := 0.5 + 0.5\sqrt{1+4/\alpha}$*,* $A_\alpha := C_\alpha + 2/\alpha$*,* $B_\alpha = \alpha C_\alpha^{\alpha+2} e^{-\frac{1}{C_\alpha}}$*.*
(iii) *For* $n = o(m)$

$$R^*_{n,m}(\mathcal{M}_0) = n\log\frac{m}{n} + \frac{3}{2}\frac{n^2}{m}\log e - \frac{3}{2}\frac{n}{m}\log e + O\left(\frac{1}{\sqrt{n}} + \frac{n^3}{m^2}\right).$$

Renewal Sources (Virtual Large Alphabet)

The **renewal process** $\mathcal{R}_0$ (introduced in 1996 by Csiszár and Shields) defined as follows:

- Let $T_1, T_2 \ldots$ be a sequence of i.i.d. positive-valued random variables with distribution $Q(j) = \Pr\{T_i = j\}$.
- 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 $T_1 - 1, T_2 - 1, \ldots$.

For a sequence

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

define k_m as the number of i such that $\alpha_i = m$. Then

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

Theorem 2 (Flajolet and W.S., 1998). *Consider the class of renewal processes. Then*

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

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

Maximal Minimax Redundancy

It can be proved that $r_{n+1} - 1 \leq D_n(\mathcal{R}_0) \leq \sum_{m=0}^{n} r_m$

$$r_n = \sum_{k=0}^{n} r_{n,k}, \quad r_{n,k} = \sum_{\mathcal{I}(n,k)} \binom{k}{k_0 \cdots k_{n-1}} \left(\frac{k_0}{k}\right)^{k_0} \left(\frac{k_1}{k}\right)^{k_1} \cdots \left(\frac{k_{n-1}}{k}\right)^{k_{n-1}}$$

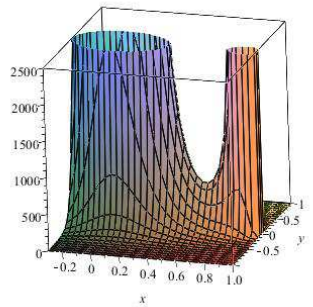
where $\mathcal{I}(n, k)$ is is the integer partition of n into k terms, i.e.,

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

But we shall study $s_n = \sum_{k=0}^{n} s_{n,k}$ where

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

since $S(z, u) = \sum_{k,n} s_{n,k} u^k z^n = \prod_{i=1}^{\infty} \beta(z^i u)$.



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

Theorem 3 (Flajolet and W.S., 1998). *We have the following asymptotics*

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

Outline

1. Workshop Objectives

2. NSF Center for Science of Information (CSoI)

3. Examples
   - Analytic Combinatorics: (Redundancy Problem of Source Coding)
   - Graphs: (**Shortest Common Superstring**)
   - Structure: (Graph Compression)

Shortest Common Superstring

Problem Formulation:
given a set of strings $X^1, X^2, \ldots, X^n$ over an alphabet $\mathcal{A}$, find the shortest string Z such that each X^i appears as a substring of Z.

Example: Consider $X^1 = abaaab$, $X^2 = aabaaaa$, $X^3 = aababb$.

X^1	=	a	b	a	a	a	b							
X^2	=				a	a	b	a	a	a	a			
X^3	=									a	a	b	a	b	b
Z	=	a	b	a	a	a	b	a	a	a	a	b	a	b	b

Observe that $\sum_i |X_i| = 19$ and $|Z| = 14$, hence the overlap $O_3 = 5$.

Optimal Overlap O_n^{opt}:
Let $\mathcal{S}$ be a set of all superstrings built over the strings $X^1, \ldots, X^n$. Then,

$$O_n^{\text{opt}} = \sum_{i=1}^{n} |X_i| - \min_{Z \in \mathcal{S}} |Z|$$

is the *optimal overlap* in the shortest common superstring.

Finding SCS is **NP-hard**.

Greedy Algorithm: On Average

Example:
Let us consider the following five strings: $X^1 = abaaab$, $X^2 = aabaaaa$, $X^3 = aababb$, $X^4 = bbaaba$, and $X^5 = bbbb$.
Let now $\mathcal{G}$ be a weighted digraph built on the set of strings $\{X^1, \ldots, X^5\}$ with weights defined as the length of the largest suffix equal to a prefix of another string.
Observe that the optimal (maximum) Hamiltonian path in $\mathcal{G}$ determines the maximum overlap between strings $X^1, \ldots, X^5$. Hence, $Z = abaaababbbbbaabaaaa$ and $O_5^{\text{opt}} = 3 + 2 + 2 + 4 = 11$.

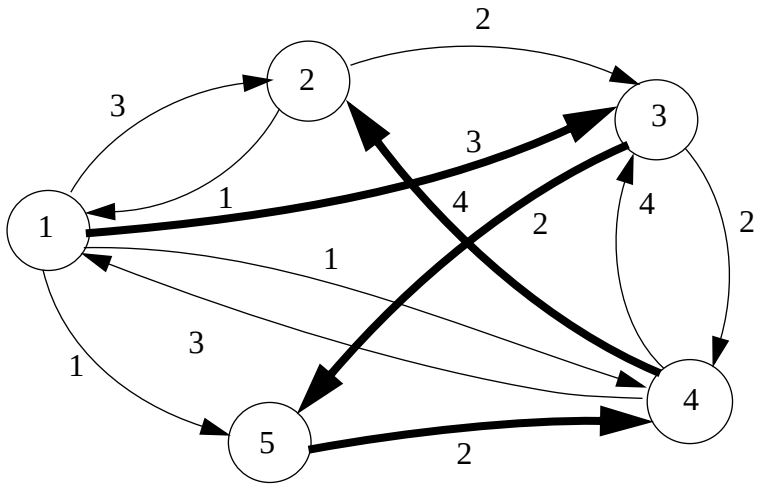


Figure 1: The digraph $\mathcal{G}$. Optimal Hamiltonian path (starting at node 4 is shown in bold.

Greedy Algorithm is Asymptotically Optimal!

Theorem 4 (Alexander, 1996; Frieze and Szpankowski, 1998). *Consider a memoryless sources that generates n independent strings over the alphabet $\mathcal{A} = \{1, \ldots, m\}$. Then*

$$\lim_{n\to\infty} \frac{O_n^{\rm opt}}{n \log n} = \frac{1}{h}, \qquad ({\rm pr.}) \qquad \lim_{n\to\infty} \frac{O_n^{\rm gr}}{n \log n} = \frac{1}{h}$$

provided the length ℓ of all strings is greater than $\frac{4}{h_1} \log n$ where $h_1 = -\ln(p_1^2 + \cdots + p_m^2)$ is the first order Rényi's entropy and $p_i = \Pr\{X^k(t) = j\}$.

Sketch of a proof. Let C_j be the longest suffix of X^1 equal to a prefix of X^j. For $O_n^{\rm opt} = \max_j C_j$ we have

$$P(\max_j C_j > t) \le nP(C_j > t) = n(p_1^2 + \cdots + p_m^2)^t.$$

This *would suggest* that whp

$$O_n^{\rm opt} \le \frac{2}{\log(p_1^2 + \cdots + p_m^2)^{-1}} \log n.$$

But it is NOT. The correct answer is whp

$$\max_j C_j = \frac{1}{h} \log n$$

where h is the entropy $h = -\sum_i p_i \log p_i$.

Outline

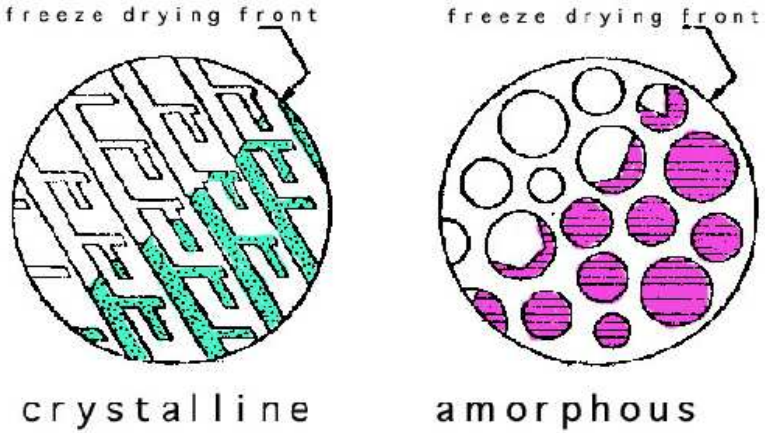
1. Workshop Objectives

2. NSF Center for Science of Information (CSoI)

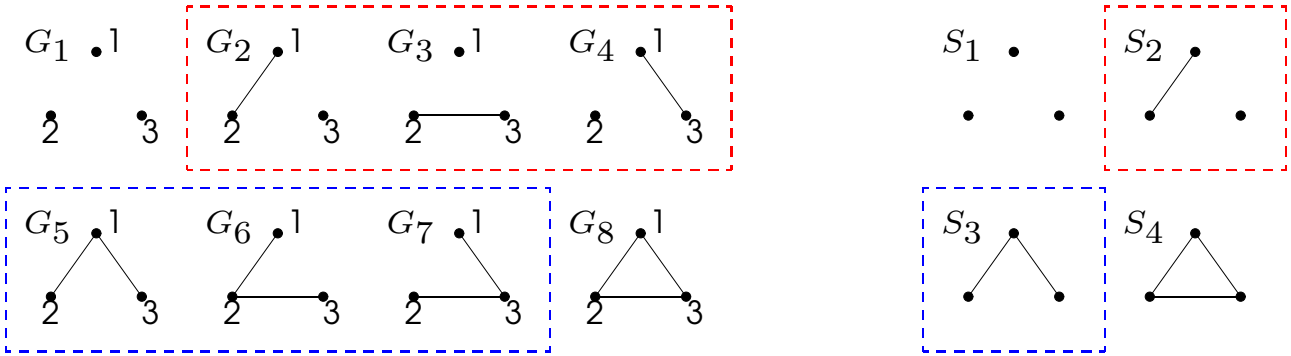
3. Examples
   - Analytic Combinatorics: (Redundancy Problem of Source Coding)
   - Graphs: (Shortest Common Superstring)
   - Structure: (**Graph Compression**)

Structure

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).



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



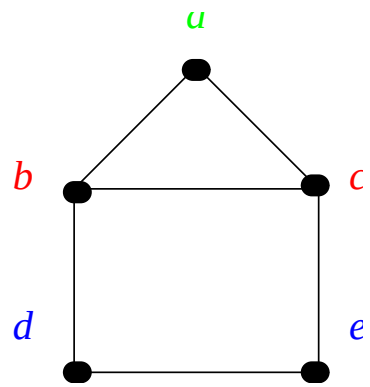
$$H_{\mathcal{G}} = \mathbf{E}[-\log P(G)] = -\sum_{G \in \mathcal{G}} P(G) \log P(G),$$

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

Automorphism and Erdös-Rényi Graph Model

Graph Automorphism:

For a graph G its automorphism is adjacency preserving permutation of vertices of G.



Erdös and Rényi model: $\mathcal{G}(n,p)$ generates graphs with n vertices, where edges are chosen independently with probability p. If G has k edges, then

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

Theorem 5 (Y. Choi and W.S., 2008). *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.),*

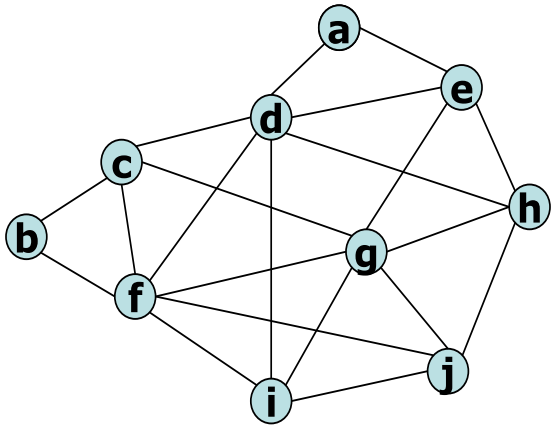
$$H_{\mathcal{S}} = \binom{n}{2} h(p) - \log n! + o(1) = \binom{n}{2} h(p) - n\log n + n\log e - \frac{1}{2}\log n + O(1),$$

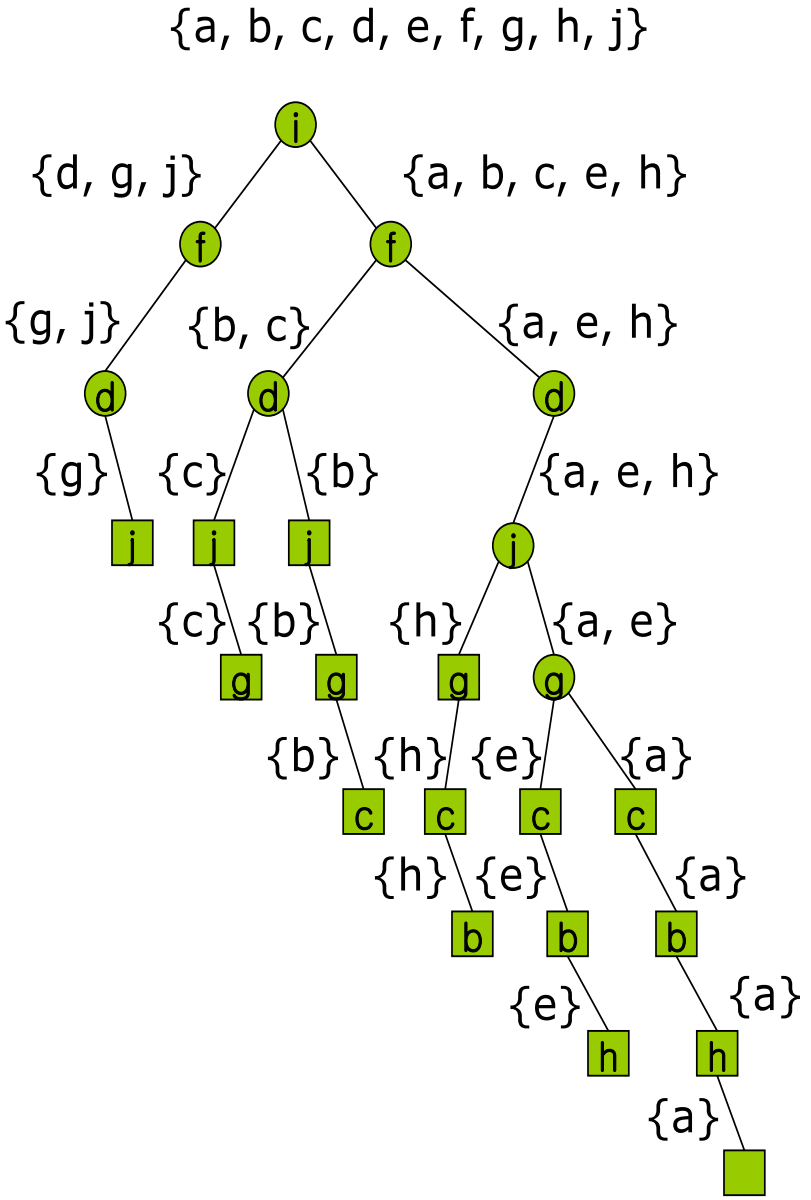
where $h(p) = -p\log p - (1-p)\log(1-p)$ is the entropy rate.

AEP *for structures:* $2^{-\binom{n}{2}(h(p)+\varepsilon)+\log n!} \le P(S) \le 2^{-\binom{n}{2}(h(p)-\varepsilon)+\log n!}$.

Sketch of Proof: **1**. $H_{\mathcal{S}} = H_{\mathcal{G}} - \log n! + \sum_{S\in\mathcal{S}} P(S)\log|\mathrm{Aut}(S)|$.
2. $\sum_{S\in\mathcal{S}} P(S)\log|\mathrm{Aut}(S)| = o(1)$ by asymmetry of $\mathcal{G}(n,p)$.

Structural Zip (SZIP) Algorithm





B1 = 01001101000011101O1

B2 = 1001011000000101

Asymptotic Optimality of SZIP

Theorem 6 (Choi, W.S., 2008). *Let $L(S)$ be the length of the code.*

(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 $h(p) = -p \log p - (1-p) \log (1-p)$, 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

That's IT