Shannon Legacy and Beyond*

Wojciech Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907

January 6, 2011

NSF CSoI

USF, Florida, 2011

*Research supported by NSF Science & Technology Center, and Humboldt Foundation.

Outline

1. Shannon Legacy

2. What is Information?

3. Post-Shannon (space, time, structure, semantics)

4. **STC on Science of Information** – NSF Science & Technology Center

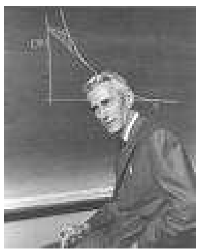
5. **Structural Information**: Graphical Compression and Fundamental Limit

Shannon Legacy

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

A Mathematical Theory of Communication.

The digital age began.



Claude Shannon:
Shannon information quantifies the extent to which a recipient of data can reduce its statistical uncertainty.
"These semantic aspects of communication are irrelevant . . ."

Applications Enabler/Driver:

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

Design Driver:

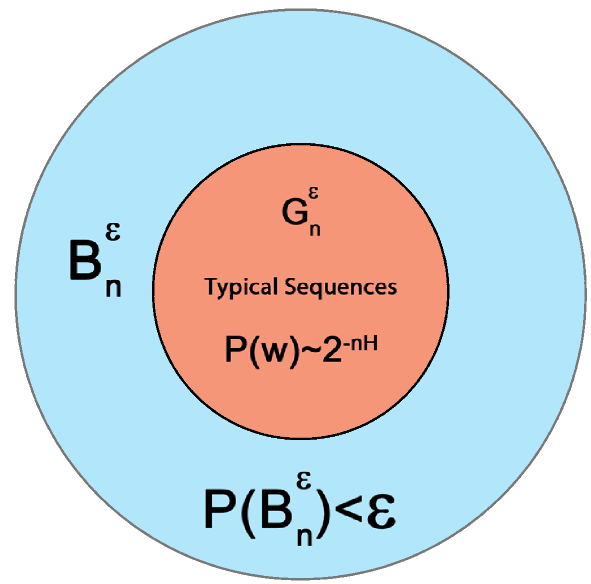
universal data compression, data encoding, voiceband modems, CDMA, multiantenna, discrete denosing, space-time codes, cryptography, . . .

Three Theorems of Shannon

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

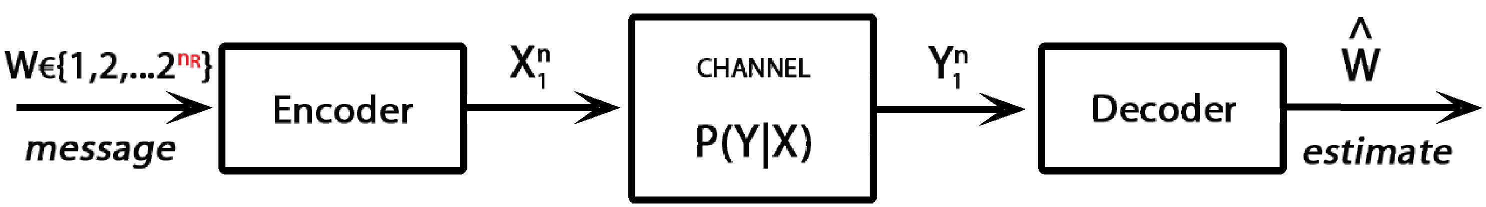
compression bit rate $\geq$ source entropy $H(X)$

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



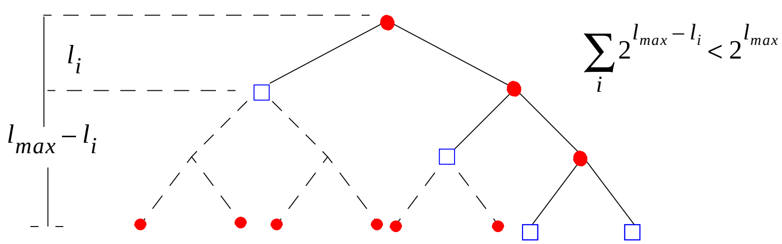
Theorem 2. (**Shannon 1948; Channel Coding**)
In Shannon's words:

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.



Theorem 1: Fundamental Limit

Prefix code is such that no codeword is a prefix of another codeword.
Kraft's Inequality: A binary code is a prefix code iff lengths $\ell_1, \ldots, \ell_N$ satisfy



$$\sum_{i=1}^{N} 2^{-\ell_i} \leq 1.$$

Shannon First Theorem: For any **prefix code** the average code length $\mathbf{E}[L(C,X)]$ cannot be smaller than the entropy $H(P)$:

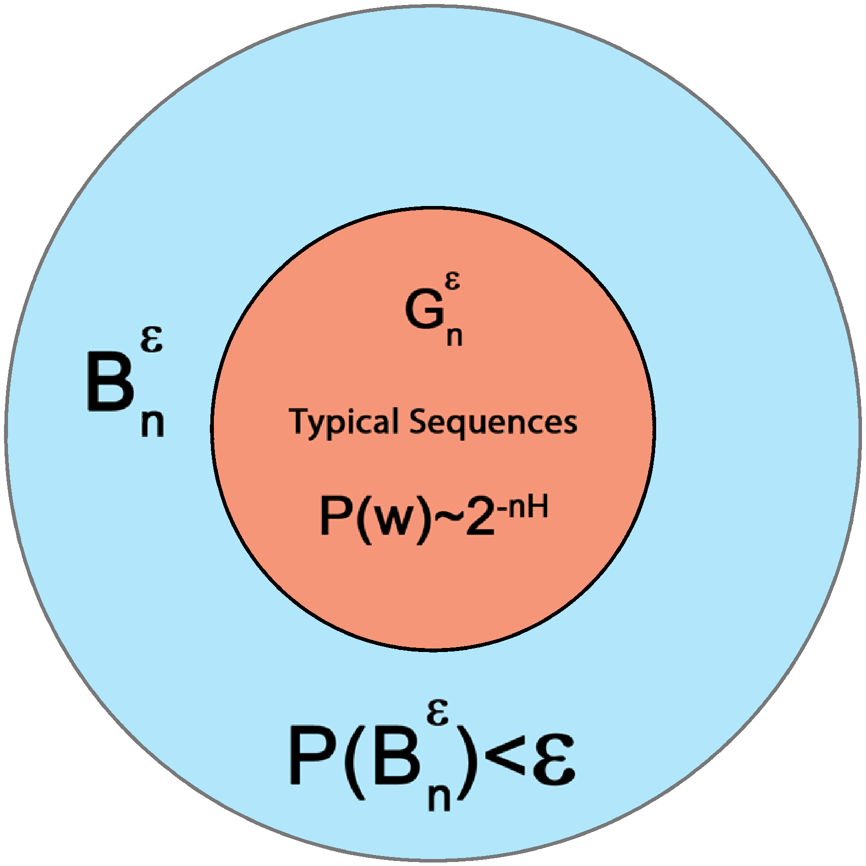
$$\mathbf{E}[L(C,X)] \geq H(P) = -\sum_{x \in \mathcal{A}^*} P(x) \log P(x).$$

Proof: Let $K = \sum_x 2^{-L(x)} \leq 1$, and $L(C,x) := L(x)$. Then

$$\begin{aligned} \mathbf{E}[L(C,X)] \quad - \quad H(P) &= \sum_{x \in \mathcal{A}^*} P(x) L(x) + \sum_{x \in \mathcal{A}^*} P(x) \log P(x) \\ &= \sum_{x \in \mathcal{A}^*} P(x) \left(-\log \frac{2^{-L(x)}/K}{P(x)} \right) - \log K \\ &\geq \sum_x P(x) - \frac{1}{K} \sum_x 2^{-L(x)} - \log K \geq 0 \end{aligned}$$

since $-\log x \geq 1 - x$ for $0 < x \leq 1$.

Theorem 1: AEP and Typical Sequences



Shannon-McMilan-Breiman:

$$-\frac{1}{n}\log P(X_1^n) \to H(X) \quad \text{(pr.)}$$

$H(X)$ is the entropy rate.

Code Length :

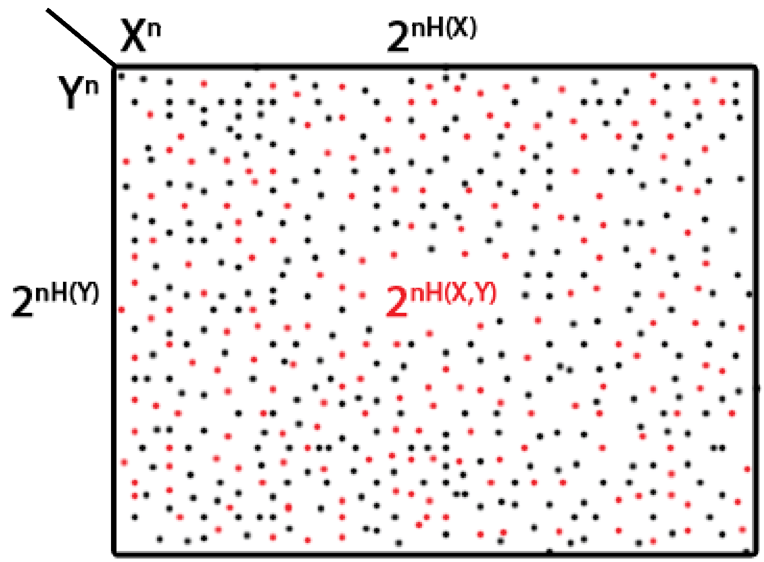
$$\lceil -\log P(X_1^n) \rceil \sim nH(X).$$

Asymptotic Equipartition Property: Sequences of length n can be partitioned into

$$\begin{array}{lll} \text{good set} & G_n^\varepsilon & P(w) \sim 2^{-nH(X)}, \quad w \in G_n^\varepsilon \\ \text{bad set} & B_n^\varepsilon & P(B_n^\varepsilon) < \varepsilon. \end{array}$$

Also, $|G_n^\varepsilon| \sim 2^{nH(X)}$.

Theorem 2: Shannon Random Decoding Rule

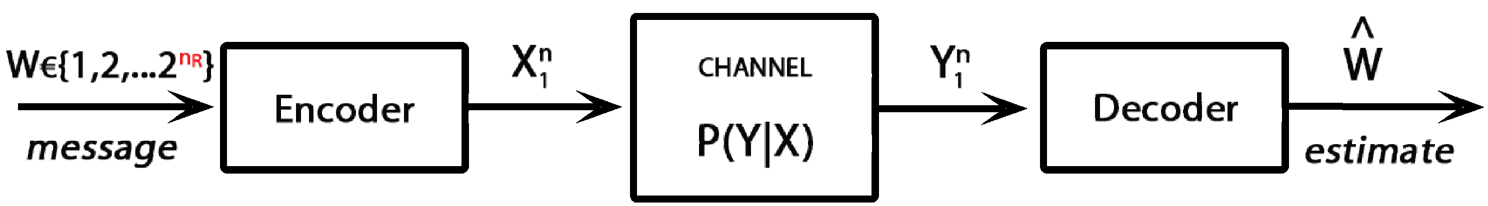


There are $2^{nH(X)}$ X-typical sequences

There are $2^{nH(Y)}$ Y-typical sequences

There are $2^{nH(X,Y)}$ jointly X,Y-typical pair of sequences

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!



Sketch of Proof: Channel Capacity Theorem

1. With high probability (whp), there is a jointly typical pair (X, Y).

2. The probability that there is another jointly typical pair is $2^{-nI(X,Y)}$:
- there are $2^{nH(X)}$ and $2^{nH(Y)}$ typical sequences X^n and Y^n, that is, $2^{n(H(X)+H(Y))}$ sequences,
- there are $2^{nH(X,Y)}$ jointly typical pairs (X, Y). The probability of error (more than one typical pair is):

$$\frac{2^{nH(X,Y)}}{2^{n(H(X)+H(Y))}} = 2^{-nI(X,Y)}.$$

Sketch of Proof: Channel Capacity Theorem

1. With high probability (whp), there is a jointly typical pair (X, Y).

2. The probability that there is another jointly typical pair is $2^{-nI(X,Y)}$:
• there are $2^{nH(X)}$ and $2^{nH(Y)}$ typical sequences X^n and Y^n, that is, $2^{n(H(X)+H(Y))}$ sequences,
• there are $2^{nH(X,Y)}$ jointly typical pairs (X, Y). The probability of error (more than one typical pair is):

$$\frac{2^{nH(X,Y)}}{2^{n(H(X)+H(Y))}} = 2^{-nI(X,Y)}.$$

3. Probability of error when 2^{nR} messages are sent is approximately

$$\min P(\mathrm{error}) \sim 2^{-n(\sup_{P(X)} I(X,Y)-R)} = 2^{-n(C-R)}.$$

Sketch of Proof: Channel Capacity Theorem

1. With high probability (whp), there is a jointly typical pair (X, Y).

2. The probability that there is another jointly typical pair is $2^{-nI(X,Y)}$:
• there are $2^{nH(X)}$ and $2^{nH(Y)}$ typical sequences X^n and Y^n, that is, $2^{n(H(X)+H(Y))}$ sequences,
• there are $2^{nH(X,Y)}$ jointly typical pairs (X, Y). The probability of error (more than one typical pair is):

$$\frac{2^{nH(X,Y)}}{2^{n(H(X)+H(Y))}} = 2^{-nI(X,Y)}.$$

3. Probability of error when 2^{nR} messages are sent is approximately

$$\min P(\text{error}) \sim 2^{-n(\sup_{P(X)} I(X,Y) - R)} = 2^{-n(C-R)}.$$

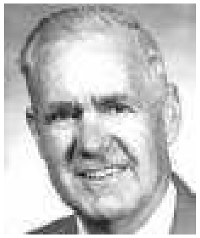
4. In conclusion:

$$\begin{array}{llll} R & < & C & P(\text{error}) \sim 2^{-n\delta} \\ R & > & C & P(\text{error}) \to 1. \end{array}$$

Outline Update

1. Shannon Legacy
2. What is Information?
3. Post-Shannon
4. STC on Science of Information
5. Structural Information

What is Information[1]?



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".

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

Event-Driven Paradigm: Systems, State, Event, Context, Attributes, Objective: Objective function objective(R, C) maps systems' rule R and context C in to an objective space.

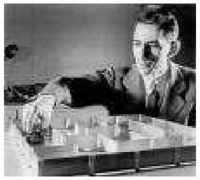
Definition 1 (J. Konorski, W.S., 2004). *The* **amount of information** *(in a faultless scenario) $I(E)$ carried by the event E in the context C as measured for a system with the rules of conduct R is*

$$I_{R,C}(E) = \mathbf{cost}[\mathrm{objective}_R(C(E)), \mathrm{objective}_R(C(E) + E)]$$

where the **cost** *(weight, distance) is a cost function.*

[1] Russell's reply to Wittgenstein's precept "whereof one cannot speak, therefore one must be silent" was "... Mr. Wittgenstein manages to say a good deal about what cannot be said."

Shannon Information



C. Shannon:
Shannon information quantifies the extent to which a recipient of data can reduce its statistical uncertainty.

Some aspects of Shannon information:

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.

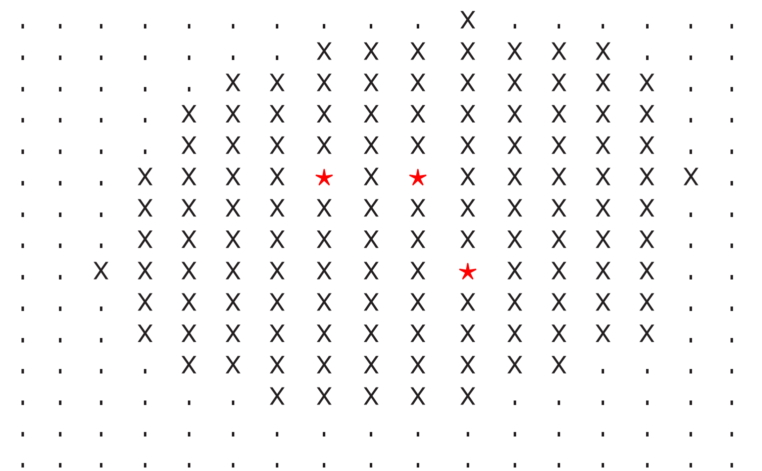
Context: "semantic aspects of communication are irrelevant"

Self-information for E_i: $I(E_i) = -\log P(E_i)$.
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).

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

Example: Distributed Information

1. In an N-threshold secret sharing scheme, N subkeys of the decryption key roam among $A \times A$ stations: (i) Event corresponds to a reception of a key; (ii) Objective(R, C) is to decode the message.



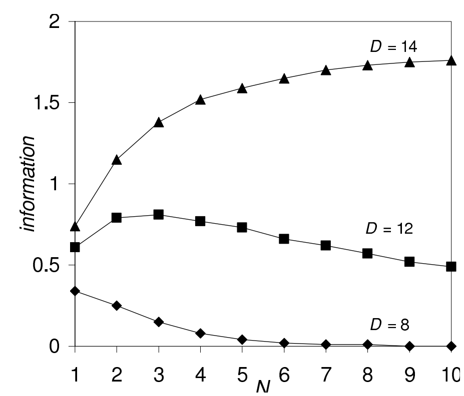
2. By protocol P a station has access:

- only it sees all N subkeys.
- it is within a distance D from all subkeys.

Note: Reception of a part of a key does not help to decrypt unless all keys are in C.

3. Assume that the larger N, the more valuable the secrets.
We define the amount of information as

I(E)= $N \times \{\# \text{ of stations having access}\}$.



Outline Update

1. Shannon Legacy
2. What is Information?
3. Post-Shannon
4. Science of Information
5. Structural Information

Post-Shannon Challenges

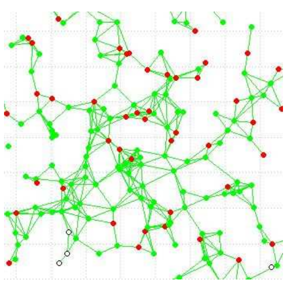
Classical Information Theory needs a recharge to meet new challenges of nowadays applications in biology, modern communication, knowledge extraction, economics and physics, ...

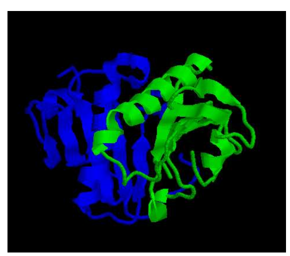
We need to extend traditional formalisms for information to include ("meaning"):

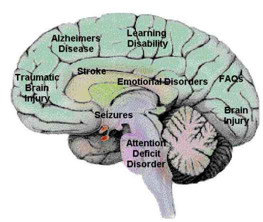
structure, time, space, and semantics,

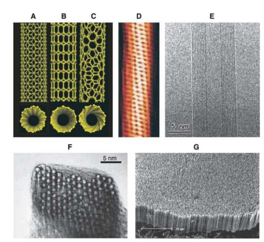
and others such as:

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



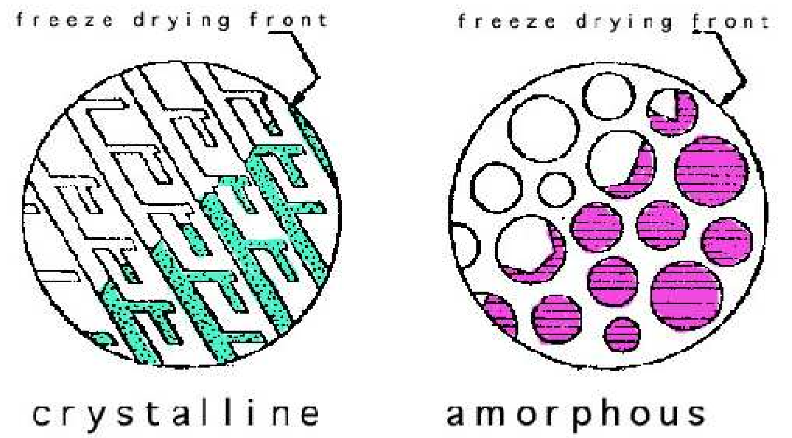




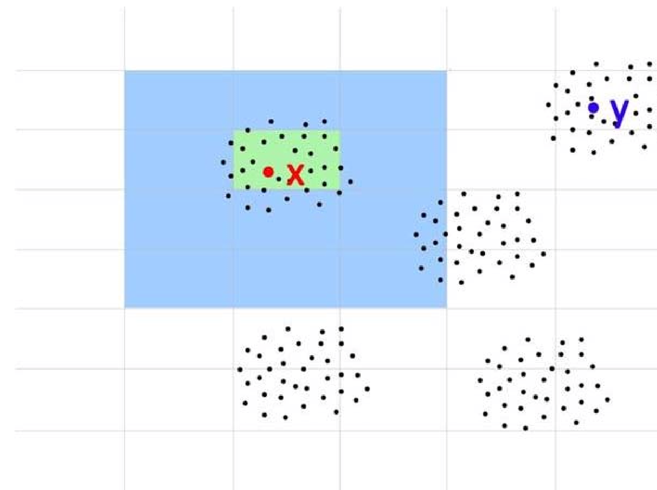


Structure, Time & Space, and Semantics

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).
(Y. Choi & W.S., ISIT, 2009.)



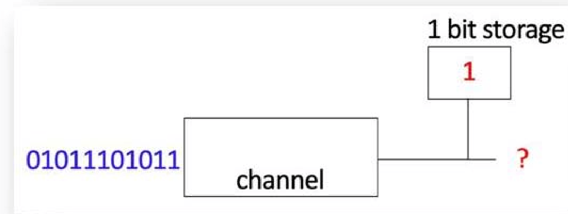
Time & Space:
Classical Information Theory is at its weakest in dealing with problems of delay
(e.g., information arriving late maybe useless or has less value).
(P. Jacquet et al., IT 2010.)



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.)

Limited Resources, Representation, and Cooperation

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?

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).

Standing on the Shoulders of Giants . . .

Manfred Eigen (Nobel Prize, 1967)
"The differentiable characteristic of the living systems is Information. Information assures the controlled reproduction of all constituents, 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**".

P. Nurse, (Nature, 2008, "Life, Logic, and Information"):
Focusing on information flow will help to understand better how cells and organisms work.

cell

". . . the generation of spatial and temporal order, memory and reproduction are not fully understood".

A. Zeilinger (Nature, 2005)
. . . reality and information are two sides of the same coin, that is, they are in a deep sense indistinguishable.

F. Brooks, jr. (JACM, 2003, "Three Great Challenges . . ."):
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.

Outline Update

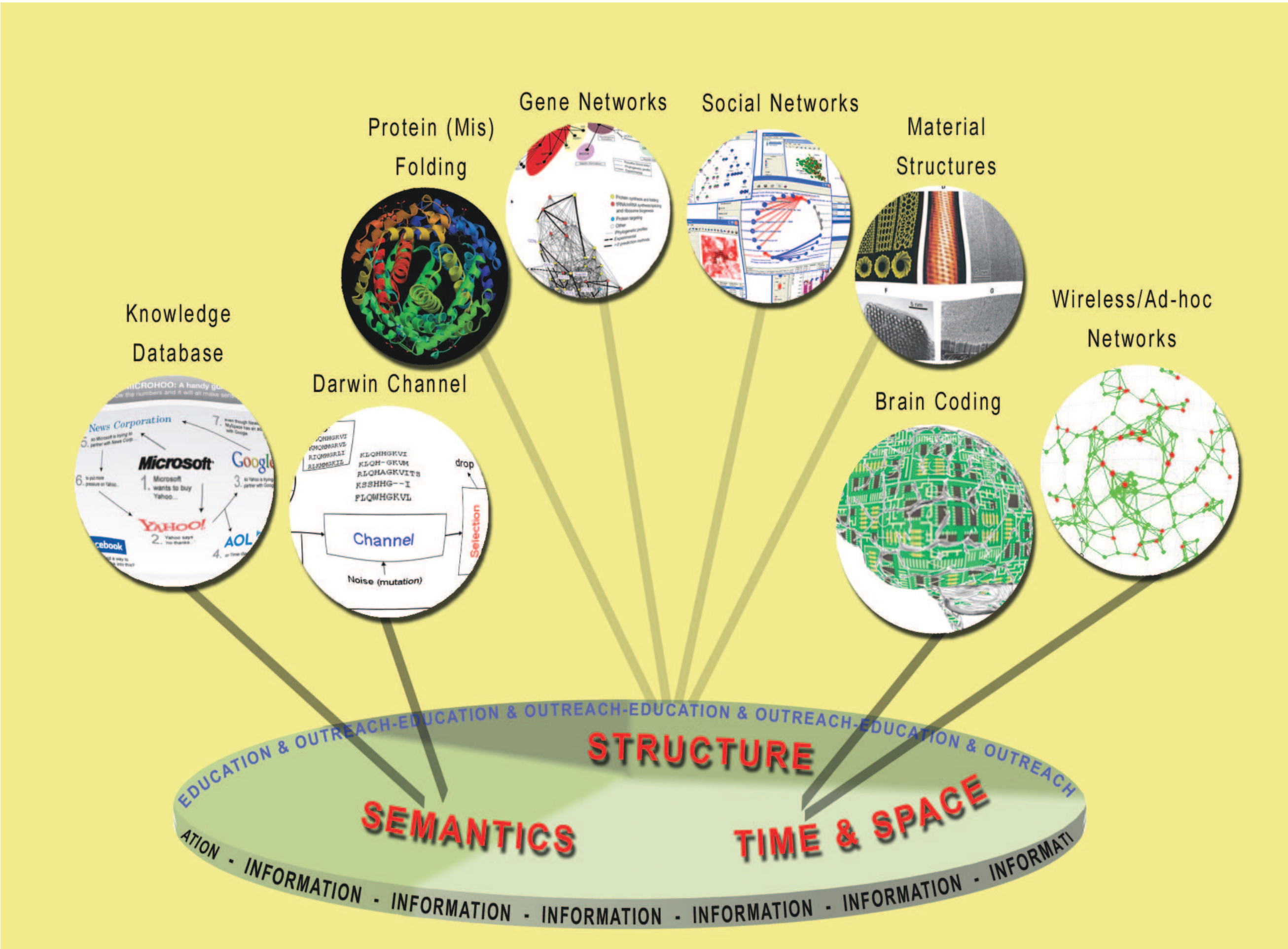
1. Shannon Legacy
2. What is Information?
3. Post-Shannon

4. NSF STC: Science of Information

5. Structural Information

Science of Information

The overarching vision of **Science of Information** is to develop rigorous principles guiding the extraction, manipulation, and exchange of information, integrating elements of space, time, structure, and semantics.



STC on Science of Information

In 2008 at Purdue we launched the

Institute for Science of Information

and in 2010 National Science Foundation established $25M

Science and Technology Center

at Purdue to do collaborative work (Berkeley, MIT, Princeton, Stanford, UIUC and Bryn Mawr & Howard U.) integrating research and teaching activities aimed at investigating the role of **information** from various viewpoints.

The specific means and goals for the Center are:

- develope post-Shannon Information Theory,
- Prestige Science Lecture Series on Information to collectively ponder short and long term goals;
- organize meetings and workshops (e.g., Information Beyond Shannon, Orlando 2005, and Venice 2008).
- initiate similar world-wide centers supporting research on information.

STC Team

Bryn Mawr College: D. Kumar

Howard University: C. Liu

MIT: M. Sudan (co-PI), P. Shor.

Purdue University (lead): W. Szpankowski (PI)

Princeton University: S. Verdu (co-PI)

Stanford University: A. Goldsmith (co-PI)

University of California, Berkeley: Bin Yu (co-PI)

University of California, San Diego: S. Subramaniam

UIUC: P.R. Kumar, O. Milenkovic.

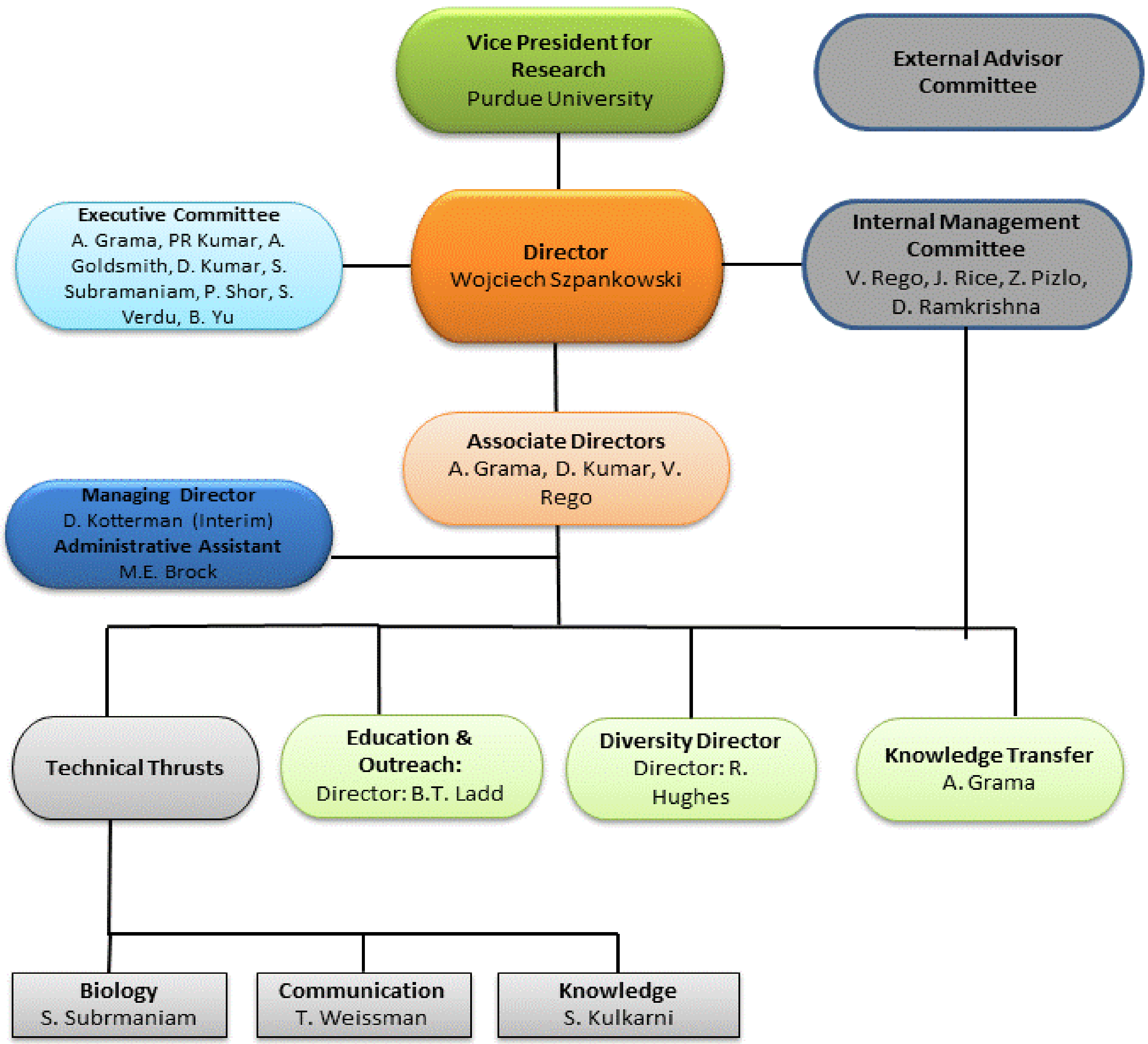
Wojciech Szpankowski, *Purdue*

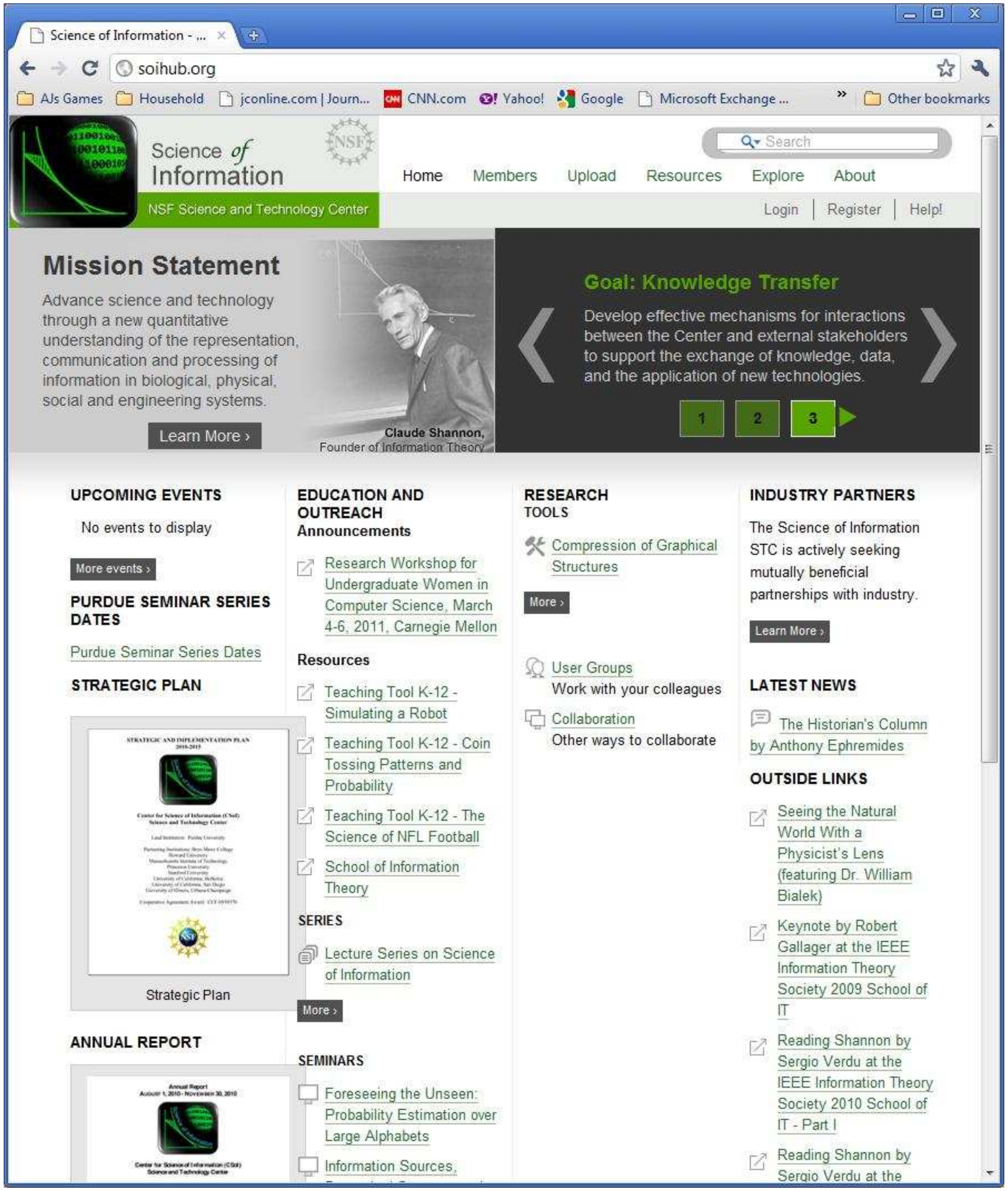
Andrea Goldsmith, *Stanford*

Madhu Sudan, *MIT*

Sergio Verdú, Princeton

Bin Yu, *U.C. Berkeley*


Vice President for Research
Purdue University
External Advisor Committee
Executive Committee
A. Grama, PR Kumar, A. Goldsmith, D. Kumar, S. Subramaniam, P. Shor, S. Verdu, B. Yu
Director
Wojciech Szpankowski
Internal Management Committee
V. Rego, J. Rice, Z. Pizlo, D. Ramkrishna
Associate Directors
A. Grama, D. Kumar, V. Rego
Managing Director
D. Kotterman (Interim)
Administrative Assistant
M.E. Brock
Technical Thrusts
Education & Outreach:
Director: B.T. Ladd
Diversity Director
Director: R. Hughes
Knowledge Transfer
A. Grama
Biology
S. Subrmaniam
Communication
T. Weissman
Knowledge
S. Kulkarni



Mission: To advance science and technology through a new quantitative understanding of the representation, communication and processing of information in biological, physical, social and engineered systems.

Outline Update

1. Shannon Legacy
2. What is Information?
3. Post-Shannon
4. STC on Science of Information

5. Structural Information: Graphical Compression and Fundamental Limit

Random Structure Model

The entropy of a random (labeled) graph process $\mathcal{G}$ is defined as

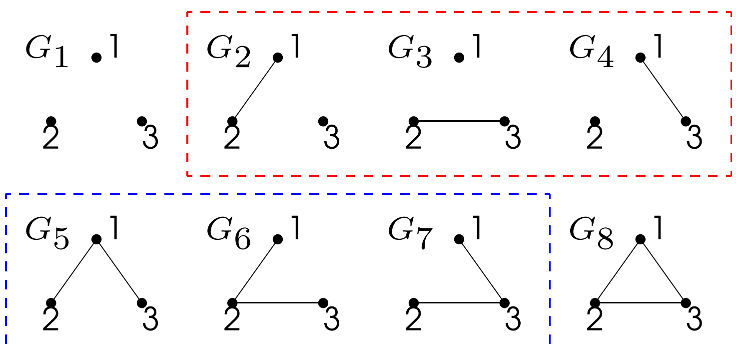
$$H_{\mathcal{G}} = \mathbf{E}[-\log P(G)] = -\sum_{G\in\mathcal{G}} P(G)\log P(G).$$

Random Structure Model

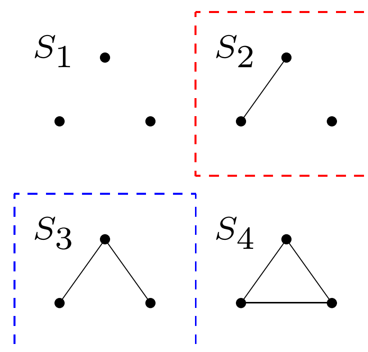
The entropy of a random (labeled) graph process $\mathcal{G}$ is defined as

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

A random structure model is defined for an unlabeled version. Some labeled graphs have the same structure.



Random Graph Model



Random Structure Model

The probability of a structure S is: $P(S) = N(S) \cdot P(G)$

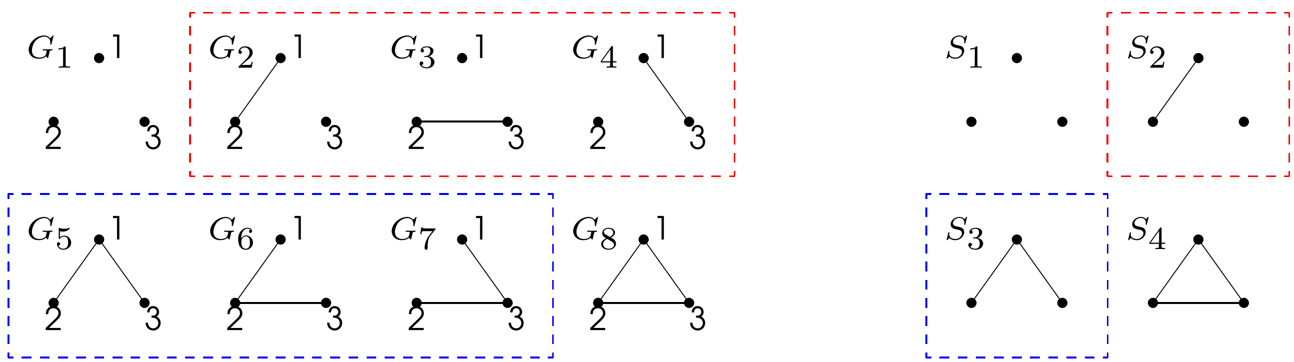
$N(S)$ is the number of different labeled graphs having the same structure.

Random Structure Model

The entropy of a random (labeled) graph process $\mathcal{G}$ is defined as

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

A random structure model is defined for an unlabeled version. Some labeled graphs have the same structure.



Random Graph Model | Random Structure Model

The probability of a structure S is: $P(S) = N(S) \cdot P(G)$

$N(S)$ is the number of different labeled graphs having the same structure.

The entropy of a random structure $\mathcal{S}$ can be defined as

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

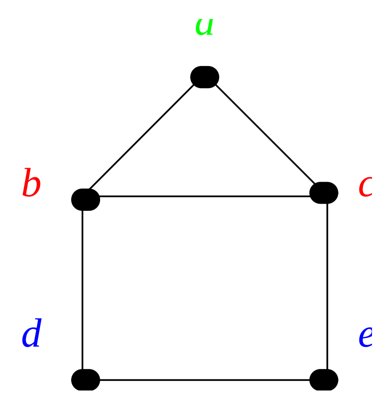
where the summation is over all distinct structures.

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

Two labeled graphs G_1 and G_2 are called *isomorphic* if and only if there is a one-to-one map from $V(G_1)$ onto $V(G_2)$ which preserves the adjacency.

Graph Automorphism:

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



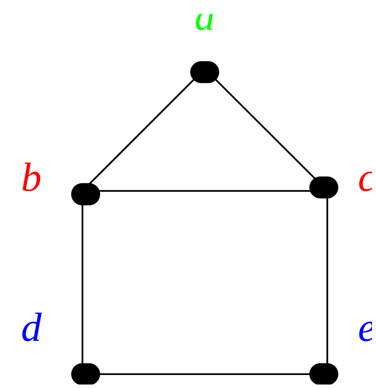
The collection $\mathrm{Aut}(G)$ of all automorphism of G is called *the automorphism group* of G.

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

Two labeled graphs G_1 and G_2 are called *isomorphic* if and only if there is a one-to-one map from $V(G_1)$ onto $V(G_2)$ which preserves the adjacency.

Graph Automorphism:

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



The collection $\mathrm{Aut}(G)$ of all automorphism of G is called *the automorphism group* of G.

Lemma 1. *If all isomorphic graphs have the same probability, then*

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

where $\mathrm{Aut}(S)$ is the automorphism group of S.

Proof idea: Using the fact that

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

Erdös-Rényi Graph Model

Our random structure model is the unlabeled version of the binomial random graph model known also as the Erdös and Rényi model.

The binomial random graph model $\mathcal{G}(n, p)$ generates graphs with n vertices, where edges are chosen independently with probability p.

If G in $\mathcal{G}(n, p)$ has k edges, then $P(G) = p^k q^{\binom{n}{2}-k}$, where $q = 1 - p$.

Erdös-Rényi Graph Model

Our random structure model is the unlabeled version of the binomial random graph model known also as the Erdös and Rényi model.

The binomial random graph model $\mathcal{G}(n, p)$ generates graphs with n vertices, where edges are chosen independently with probability p.

If G in $\mathcal{G}(n, p)$ has k edges, then $P(G) = p^k q^{\binom{n}{2}-k}$, where $q = 1 - p$.

Theorem 1 (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:* $\quad 2^{-\binom{n}{2}(h(p)+\varepsilon)+\log n!} \leq P(S) \leq 2^{-\binom{n}{2}(h(p)-\varepsilon)+\log n!}.$

Erdös-Rényi Graph Model

Our random structure model is the unlabeled version of the binomial random graph model known also as the Erdös and Rényi model.

The binomial random graph model $\mathcal{G}(n, p)$ generates graphs with n vertices, where edges are chosen independently with probability p.

If G in $\mathcal{G}(n, p)$ has k edges, then $P(G) = p^k q^{\binom{n}{2}-k}$, where $q = 1 - p$.

Theorem 1 (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:* $\quad 2^{-\binom{n}{2}(h(p)+\varepsilon)+\log n!} \leq P(S) \leq 2^{-\binom{n}{2}(h(p)-\varepsilon)+\log n!}.$

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

Compression Algorithm

Compression Algorithm called Structural zip, in short SZIP – Demo.

Compression Algorithm

Compression Algorithm called Structural zip, in short SZIP – Demo.

We can prove the following estimate on the compression ratio of $S(p,n)$ for our algorithm SZIP.

Theorem 2. *Let* $L(S)$ *be the length of the code generated by our algorithm for all graphs* G *from* $\mathcal{G}(n,p)$ *that are isomorphic to a structure* S.
(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 $h(p) = -p\log p - (1-p)\log(1-p)$, 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) *Finally, our algorithm runs in* $O(n+e)$ *on average, where* e *is the number of edges.*

Our algorithm is asymptotically optimal up to the second largest term, and works quite fine in practise.

Experimental Results

Real-world and random graphs.

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	37,018,710	729,848	310,569
	Internet (AS level)	25,881	52,407	301,148	334,900,140	1,572, 210	396,060
Random	$\mathcal{S}(n,p)$	1,000	$p=0.01$	34,361	499,500	99,900	40,350
	$\mathcal{S}(n,p)$	1,000	$p=0.1$	227,236	499,500	999,999	234,392
	$\mathcal{S}(n,p)$	1,000	$p=0.3$	432,692	499,500	2,997,99 9	440,252
					$\binom{n}{2}$	$2e\lceil\log n\rceil$	$\binom{n}{2}h(p)$

- n : number of vertices
- e : number of edges
- Adjacency matrix : $\binom{n}{2}$ bits
- Adjacency list : $2e\lceil\log n\rceil$ bits
- Arithmetic coding : $\sim \binom{n}{2}h(p)$ bits (compressing the adjacency matrix)

That's It

THANK YOU