Facets of Information*

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

May 16, 2010

AofA and IT logos

Stuttgart, 2010

*Research supported by NSF Science & Technology Center, and Humboldt Foundation.
†Joint work with Y. Choi, Purdue, P. Jacquet, INRIA, France.
1. Shannon Legacy
2. Post-Shannon (space, time, structure, semantics)
3. Science of Information
4. Learnable Information – Fundamental Limit of Information Extraction
5. Transfer of Spatio-Temporal Information – Speed of Information
6. Structural Information: Graphical Compression and Fundamental Limit
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, . . .
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).

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

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

- **Compression Bound:**
  \[
  \text{bit rate} \geq H(X)
  \]
  for distortion level \( D \):
  - **Lossy Bit Rate:**
    \[
    \text{lossy bit rate} \geq R(D)
    \]

**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.
Outline Update

1. Shannon Legacy
2. Post-Shannon
3. Science of Information
4. Learnable Information – Fundamental Limits
5. Spatio-Temporal Information
6. 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:
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).

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).

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?
Limited Computational Resources:
In many scenarios, information is limited by available computational resources (e.g., cell phone, living cell).

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.)
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”.

Focusing on information flow will help to understand better how cells and organisms work.
“. . . 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.

C. F. Von Weizsäcker: What is Information\(^1\)?
“Information is only that which produces information” (relativity).
“Information is only that which is understood” (rationality)
“Information has no absolute meaning”.
Outline Update

1. Shannon Legacy
2. Post-Shannon
3. NSF STC: Science of Information
4. Learnable Information – Fundamental Limits
5. Spatio-Temporal Information
6. Structural 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.
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 with 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: from the fundamental theoretical underpinnings of green information to the science and engineering of novel information substrates, biological pathways, communication networks, economics, and complex social systems.

The specific means and goals for the Center are:

- develop 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.
Outline Update

1. Shannon Legacy
2. Post-Shannon
3. Science of Information
4. Learnable Information – Fundamental Limits
5. Spatio-Temporal Information
6. Structural Information
7. Science of Information
How much **useful information** can be **extracted** from a data set?

**Eugene Wigner:**
Physics doesn’t describe **nature**.
Physics describes **regularities** among events . . .
Learnable Information

How much useful information can be extracted from a data set?

Eugene Wigner:
Physics doesn’t describe nature.
Physics describes regularities among events . . .

1. Let $S$ be a set representing regularity or summarizing properties of a sequence $x^n = x_1 \ldots x_n$ of length $n$ (e.g., $S$ is the number of 1 in $x^n$).

2. We can represent $x^n$ by:
(i) describing the set $S$ that takes $I(S)$ bits – we call it useful information;
(ii) position of $x^n$ in $S$ that requires $\log |S|$ bits (complexity of $x^n$).

3. Kolmogorov Information: Define

$$K(x^n) = K(\hat{S}) + \log |\hat{S}|.$$  

Example: For $x^n$ being a binary sequence, let $S$ be the type of $x^n$ that requires $K(\hat{S}) = \frac{1}{2} \log n$ bits, and

$$\log |S| = \log \binom{n}{k} = nH(n/k)$$ bits.
1. $\mathcal{M}_k = \{P_\theta : \theta \in \Theta\}$ set of $k$-dimensional parameterized distributions. Let $\hat{\theta}(x^n) = \arg\max_{\theta \in \Theta} [\log 1 / P_\theta(x^n)]$ be the ML estimator. (demo)
Computable Learnable Information (Rissanen)

1. $\mathcal{M}_k = \{P_\theta : \theta \in \Theta\}$ set of $k$-dimensional parameterized distributions. Let $\hat{\theta}(x^n) = \arg \max_{\theta \in \Theta} [\log 1/P_\theta(x^n)]$ be the ML estimator. (demo)

2. Two models, say $P_\theta(x^n)$ and $P_{\theta'}(x^n)$ are indistinguishable (have the same useful information) if the ML estimator $\hat{\theta}$ with high probability declares both models are the same (i.e., $\theta$ and $\theta'$ are close).

3. The number of distinguishable distributions (i.e, $\hat{\theta}$) $C_n(\Theta)$ summarizes then the learnable information and we denote it as $I(\Theta) = \log_2 C_n(\Theta)$. 

1. $\mathcal{M}_k = \{P_\theta : \theta \in \Theta\}$ set of $k$-dimensional parameterized distributions. Let $\hat{\theta}(x^n) = \arg \max_{\theta \in \Theta} [\log 1/P_\theta(x^n)]$ be the ML estimator. (demo)

2. Two models, say $P_\theta(x^n)$ and $P_{\theta'}(x^n)$ are indistinguishable (have the same useful information) if the ML estimator $\hat{\theta}$ with high probability declares both models are the same (i.e., $\theta$ and $\theta'$ are close).

3. The number of distinguishable distributions (i.e, $\hat{\theta}$) $C_n(\Theta)$ summarizes then the learnable information and we denote it as $I(\Theta) = \log_2 C_n(\Theta)$.

4. Balasubramanian proved that the number of distinguishable balls $C_n(\Theta)$ of radius $O(1/\sqrt{n})$ is equal to (i.e., minimax maximal regret)

$$I(\Theta) = \log C_n(\Theta) \sim \inf_{\theta \in \Theta} \max_{x^n} \log \frac{P_{\theta}}{P_\theta} = \log \sum_{x^n} P_{\hat{\theta}}(x^n) = \text{minimum regret}.$$
1. $\mathcal{M}_k = \{P_\theta : \theta \in \Theta\}$ set of $k$-dimensional parameterized distributions. Let $\hat{\theta}(x^n) = \arg \max_{\theta \in \Theta} [\log 1/ P_\theta(x^n)]$ be the ML estimator. (demo)

2. Two models, say $P_\theta(x^n)$ and $P_{\theta'}(x^n)$ are indistinguishable (have the same useful information) if the ML estimator $\hat{\theta}$ with high probability declares both models are the same (i.e., $\theta$ and $\theta'$ are close).

3. The number of distinguishable distributions (i.e., $\hat{\theta}$) $C_n(\Theta)$ summarizes then the learnable information and we denote it as $I(\Theta) = \log_2 C_n(\Theta)$.

4. Balasubramanian proved that the number of distinguishable balls $C_n(\Theta)$ of radius $O(1/\sqrt{n})$ is equal to (i.e., minimax maximal regret)

$$I(\Theta) = \log C_n(\Theta) \sim \inf_{\theta \in \Theta} \max_{x^n} \log \frac{P_{\hat{\theta}}}{P_\theta} = \log \sum_{x^n} P_{\hat{\theta}}(x^n) = \text{minimum regret}.$$
Consider the **minimax regret**=useful information for memoryless sources of \(m\)-ary alphabet, thus with \(k = m - 1\). Observe that

\[
C_n(\Theta) = \sum_{x_1^n} \sup_{p_1, \ldots, p_m} p_1^{k_1} \cdots p_m^{k_m} = \sum_{k_1 + \cdots + k_m = n} \binom{n}{k_1, \ldots, k_m} \left( \frac{k_1}{n} \right)^{k_1} \cdots \left( \frac{k_m}{n} \right)^{k_m}.
\]

which becomes: \(C_n(\Theta) = \frac{n!}{n^n} \sum_{k_1 + \cdots + k_m = n} \frac{k_1^{k_1}}{k_1!} \cdots \frac{k_m^{k_m}}{k_m!} \).

Using **tree-generating functions**

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

where \(T(z)\) satisfies \(T(z) = z e^{T(z)}\) we conclude that \(D_m(z) = \sum_{n=0}^{\infty} \frac{n^n}{n!} z^n C_n(\Theta)\) becomes

\[
D_m(z) = \left[ B(z) \right]^m.
\]

The singularity analysis yields

\[
I_n(\Theta) = \frac{m - 1}{2} \log \left( \frac{n}{2} \right) + \log \left( \frac{\sqrt{\pi}}{\Gamma\left(\frac{m}{2}\right)} \right) + \frac{\Gamma\left(\frac{m}{2}\right) m}{3 \Gamma\left(\frac{m}{2} - \frac{1}{2}\right)} \cdot \frac{\sqrt{2}}{\sqrt{n}} + \ldots
\]
Outline Update

1. Shannon Legacy
2. Post-Shannon
3. Learnable Information – Fundamental Limits
4. Spatio-Temporal Information – Speed of Information
5. Structural Information
Transfer of Information in Ubiquitous Networks

Information Theory, born 50 years ago, needs a recharge if it is to meet new challenges of ubiquitous networks.

Fundamental New Problems:

1. **Future networks** will transport information, not data.

2. **Information** is only useful when delivered in a timely fashion (e.g., needs new resource scheduling in inherently unreliable wireless environment).

3. To design scalable networks, node must cooperate (e.g., interference can be turned into useful signals through distributive multiantenna processing; mobility may diffuse traffic but will cause large delays).

4. New **Information Theory of dependence** is needed to design more energy efficient communication (i.e., how fast? at what cost?).

5. To turn all of these into reality, we must seriously consider:
   - **selfishness** (it is in each node’s self-interest to cooperate);
   - **channel capacity** (to turn interference into useful signals);
   - **delay** (mobility can diffuse traffic for large delays).
**Speed of Information**

Based on P. Jacquet, B. Mans and G. Rodolakis, ISIT, 2008

**Intermittently Connected Mobile Networks (ICN):**

1. Nodes move in space with uniform density $\nu > 0$.
2. Nodes do random walks with speed $v$ and turn rate $\tau$.
3. Connectivity is achieved in a unit disk.
4. Radio propagation speed is infinite.

**Problem statement:**
At time $t = 0$ a node at the origin broadcasts a beacon and nodes retransmit beacon immediately to neighbors in the ICN network.

**Question:** At what time $T$ node at distance $L$ from the origin will receive the beacon? **Propagation speed** is $\frac{L}{T}$. 
Journey Analysis Through the Laplace Transform

Let \( p(z, t) \) be the space-time density of (journey) paths \( C \) that reaches location \( z(C) \) at time \( t \).

(Probabilistic) information speed is the smallest \( \sigma_0 \) such that for all \( \sigma > \sigma_0 \)
\[
\lim_{\sigma \to \sigma_0} p \left( z, \frac{|z|}{\sigma} \right) = 0.
\]

Example: For \( p(z, t) = O(\exp(-A|z| + Bt + C)) \), then \( \sigma_0 = B/A \).

Theorem 1 (Jacquet, et. al., 2008). The information speed is not greater than the smallest ratio \( \frac{\theta}{\rho} \) where \((\rho, \theta)\) are roots of a function \( D(\rho, \theta) = 0 \).

![Figure 1: Time versus distance and Impact on the network capacity.](image)

- Mobility creates capacity

Information propagation time

Permanently disconnected

\( T(z') \)

Permanently connected
Outline Update

1. Shannon Legacy
2. Post-Shannon
3. Learnable Information – Fundamental Limits
4. Spatio-Temporal Information
5. Structural Information: Graphical Compression and Fundamental Limit
The entropy of a random (labeled) graph process $G$ is defined as

$$H_G = \mathbb{E}[-\log P(G)] = -\sum_{G \in G} P(G) \log P(G).$$
The entropy of a random (labeled) graph process $G$ is defined as

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

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

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 (labeled) graph process $G$ is defined as

$$H_G = \mathbb{E}[ - \log P(G)] = - \sum_{G \in G} P(G) \log P(G).$$

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

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 $S$ can be defined as

$$H_S = \mathbb{E}[ - \log P(S)] = - \sum_{S \in S} P(S) \log P(S),$$

where the summation is over all distinct structures.
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 $\text{Aut}(G)$ of all automorphism of $G$ is called the *automorphism group* of $G$.
Relationship between $H_G$ and $H_S$

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 $\text{Aut}(G)$ of all automorphism of $G$ is called the automorphism group of $G$.

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

$$H_S = H_G - \log n! + \sum_{S \in \mathcal{S}} P(S) \log |\text{Aut}(S)|,$$

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

Proof idea: Using the fact that

$$N(S) = \frac{n!}{|\text{Aut}(S)|}.$$
Our random structure model is the unlabeled version of the binomial random graph model known also as the Erdős and Rényi model.

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

If $G$ in $G(n, p)$ has $k$ edges, then $P(G) = p^k q^{n^2 - k}$, 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 $G(n, p)$ generates graphs with $n$ vertices, where edges are chosen independently with probability $p$.

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

**Theorem 2 (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.),

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

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

**AEP for structures:**

$$2^{-\left(\binom{n}{2} (h(p) + \varepsilon) + \log n!\right)} \leq P(S) \leq 2^{-\left(\binom{n}{2} (h(p) - \varepsilon) + \log n!\right)}.$$
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 $G(n, p)$ generates graphs with $n$ vertices, where edges are chosen independently with probability $p$.

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

Theorem 2 (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.),

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

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

AEP for structures:

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

Proof idea: 1. $H_S = H_G - \log n! + \sum_{S \in S} P(S) \log |\text{Aut}(S)|$.
2. $H_G = \binom{n}{2} h(p)$
3. $\sum_{S \in S} P(S) \log |\text{Aut}(S)| = o(1)$ by asymmetry of $G(n, p)$. 
Compression Algorithm called Structural zip, in short SZIP – Demo.
Compression Algorithm

Compression Algorithm called Structural zip, in short \texttt{SZIP} – Demo.

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

\textbf{Theorem 3.} Let $L(S)$ be the length of the code generated by our algorithm for all graphs $G$ from $G(n, p)$ that are isomorphic to a structure $S$.

(i) For large $n$,

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

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 $\varepsilon > 0$,

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

(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 practice.
Experimental Results

Real-world and random graphs.

Table 1: The length of encodings (in bits)

<table>
<thead>
<tr>
<th>Networks</th>
<th># of nodes</th>
<th># of edges</th>
<th>our algorithm</th>
<th>adjacency matrix</th>
<th>adjacency list</th>
<th>arithmetic coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>US Airports</td>
<td>332</td>
<td>2,126</td>
<td>8,118</td>
<td>54,946</td>
<td>38,268</td>
<td>12,991</td>
</tr>
<tr>
<td>Protein interaction (Yeast)</td>
<td>2,361</td>
<td>6,646</td>
<td>46,912</td>
<td>2,785,980</td>
<td>159,504</td>
<td>67,488</td>
</tr>
<tr>
<td>Collaboration (Geometry)</td>
<td>6,167</td>
<td>21,535</td>
<td>115,365</td>
<td>19,012,861</td>
<td>55,9,910</td>
<td>241,811</td>
</tr>
<tr>
<td>Collaboration (Erdős)</td>
<td>6,935</td>
<td>11,857</td>
<td>62,617</td>
<td>24,043,645</td>
<td>308,2,82</td>
<td>147,377</td>
</tr>
<tr>
<td>Genetic interaction (Human)</td>
<td>8,605</td>
<td>26,066</td>
<td>221,199</td>
<td>37,018,710</td>
<td>729,848</td>
<td>310,569</td>
</tr>
<tr>
<td>Internet (AS level)</td>
<td>25,881</td>
<td>52,407</td>
<td>301,148</td>
<td>334,900,140</td>
<td>1,572,210</td>
<td>396,060</td>
</tr>
<tr>
<td>$S(n, p)$</td>
<td>1,000</td>
<td>$p = 0.01$</td>
<td>34,361</td>
<td>499,500</td>
<td>99,900</td>
<td>40,350</td>
</tr>
<tr>
<td>$S(n, p)$</td>
<td>1,000</td>
<td>$p = 0.1$</td>
<td>227,236</td>
<td>499,500</td>
<td>999,999</td>
<td>234,392</td>
</tr>
<tr>
<td>$S(n, p)$</td>
<td>1,000</td>
<td>$p = 0.3$</td>
<td>432,692</td>
<td>499,500</td>
<td>2,997,999</td>
<td>440,252</td>
</tr>
</tbody>
</table>

- $n$: 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
Both quantities $E[|B_1|]$ and $E[|B_2|]$ can be analyzed by solving the following two-dimensional recurrence:

\[
a_{n+1,0} = c_n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (a_{k,0} + a_{n-k,k}),
\]

\[
a_{n,d} = c_n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (a_{k,d-1} + a_{n-k,k+d-1}).
\]

for some given $c_n$ (e.g., $c_n = \lceil \log(n+1) \rceil$ for $E[|B_1|]$).

Similar (one-dimensional) recurrences appear in tries and digital search trees.

**Trie** : \(x_n = c_n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (x_k + x_{n-k}).\)

**DST** : \(y_{n+1} = c_n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (y_k + y_{n-k}).\)

We can show that

\[y_n \leq a_{n,d} \leq x_n.\]
Theorem 4. For large $n$, 

$$
\mathbb{E}[|B_1|] \leq \frac{1}{h} (\beta + \Phi_1(\log n)) n + o(n),
$$

where $h := h(p)$,

$$
\beta = \log e \cdot \sum_{b \geq 2} \frac{[\log (b + 1)]}{b(b - 1)} = 3.760 \cdots,
$$

and $\Phi_1(\log n)$ is a fluctuating function for $\log p/\log q$ rational with small amplitude and zero otherwise.

Proof idea: First, we can prove that 

$$
\mathbb{E}[|B_1|] \leq x_n
$$

where $x_n$ satisfies $x_0 = x_1 = 0$ and for $n \geq 2$

$$
x_n = [\log (n + 1)] + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (x_k + x_{n-k}).
$$

This recurrence can be solved using analytic techniques such as generating functions, Mellin transform, and Poissonization.
Theorem 5. For large $n$,

$$
\mathbb{E}[|B_2|] \leq \frac{n(n - 1)}{2} - \frac{n}{h} \log n
$$

$$
+ \frac{n}{h} \left( \frac{h}{2} - \frac{h_2}{2h} - \gamma + 1 + \alpha - \Phi_2(\log n) \right) - \frac{1}{h} \log n + O(1),
$$

where $h := h(p)$, $\gamma = 0.577 \cdots$ is the Euler constant, $h_2 = p \log^2 p + q \log^2 q$, $\alpha = -\sum_{k=1}^{\infty} \frac{p^{k+1} \log p + q^{k+1} \log q}{1 - p^{k+1} - q^{k+1}}$, and $\Phi_2(\log n)$ is a fluctuating function for $\log p/\log q$ rational with small amplitude and zero otherwise.

Proof idea: We can first prove that

$$
\mathbb{E}[|B_2|] = \frac{n(n - 1)}{2} - b_n
$$

for some $b_n$. Observe that $b_n \geq y_n - \frac{n}{2}$ for some $y_n$ satisfying $y_0 = 0$ and $y_n$ is the expected path length in a digital search tree for $n \geq 0$ (i.e., $y_n$ is the expected path length in a digital search tree)

$$
y_{n+1} = n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (y_k + y_{n-k}).
$$
THANK YOU