Structural Information

Wojciech Szpankowski
Purdue University
W. Lafayette, IN 47907

April 26, 2014

Center for Science of Information
NSF Science & Technology Center

LIDS, MIT, Boston, 2014
Information Theory of Data Structures: Following Ziv (1997) we propose to explore finite size information theory of data structures (i.e., sequences, graphs), that is, to develop information theory of various data structures beyond first-order asymptotics. We focus here on information of graphical structures (unlabeled graphs).

F. Brooks, jr. (JACM, 50, 2003, “Three Great Challenges for ... CS“):

“We have no theory that gives us a metric for the Information embodied in structure. This is the most fundamental gap in the theoretical underpinnings of information science and of computer science.”

Networks (Internet, protein-protein interactions, and collaboration network) and Matter (chemicals and proteins) have structures. They can be abstracted by (unlabeled) graphs.
Outline

1. Structural Compression
   - Motivation
   - Unlabeled Graphs
   - SZIP Algorithm and Its Analysis
   - Structural Binary Symmetric Channel

2. Structure of Markov Fields
   - Markov Types
   - One-Dimensional Markov Chains
   - One-Dimensional Universal Types
   - Markov Fields and Tilings

3. Sequence-Structure Protein Folding Channel
Graphs with Locally Correlated Labels

How many bits are required to describe the unlabelled graph on the left, and how many additional bits one needs to represent the correlated labels on the right?
Figure 1: Protein-Protein Interaction Network with BioGRID database
1. **Structural Compression**
   - Motivation
   - Unlabeled Graphs
   - SZIP Algorithm and Its Analysis
   - Structural Binary Symmetric Channel

2. Structure of Markov Fields

3. Sequence-Structure Protein Folding Channel
Information Content of Unlabeled Graphs:

A structure model $S$ of a graph $G$ 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.

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

$$H_S = \mathbb{E}[- \log P(S)] = - \sum_{S \in S} P(S) \log P(S), \quad \text{structural entropy}$$
Two labeled graphs $G_1$ and $G_2$ are called isomorphic if and only if there is a one-to-one mapping 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

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

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

**Proof idea:** Using the fact that

$$N(S) = \frac{n!}{|\text{Aut}(S)|}.$$
Erdős-Rényi Graph Model

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

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

If a graph $G$ in $G(n, p)$ has $k$ edges, then (where $q = 1 - p$)

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

Lemma 2 (Kim, Sudakov, and Vu, 2002). For Erdős-Rényi graphs and all $p$ satisfying

$$\frac{\ln n}{n} \ll p, \quad 1 - p \gg \frac{\ln n}{n}$$

a random graph $G \in G(n, p)$ is symmetric (i.e., $\text{Aut}(G) \approx 1$) with probability $O\left(n^{-w}\right)$ for any positive constant $w$, that is,

$$P(\text{Aut}(G) = 1) \sim 1 - O(n^{-w}).$$
Symmetry of Power Law Graphs?

Figure 2: Logarithm of the number of automorphisms versus the number of vertices for $m = 1$ (on the left), $m = 4$ (middle), defect for various $m$.
Symmetry of Power Law Graphs?

Figure 2: Logarithm of the number of automorphisms versus the number of vertices for $m = 1$ (on the left), $m = 4$ (middle), defect for various $m$.

**Theorem 1** (Symmetry Results for $m = 1, 2$). Let graph $G_n$ be generated by the preferential model with parameter $m = 1$ or $m = 2$. Then there exists a constant $C > 0$ such that, for $n$ sufficiently large,

$$\Pr[|\text{Aut}(G_n)| > 1] > C.$$ 

**Conjecture 1.** For $m \geq 3$ a graph $G_n$ generated by the preferential model is asymmetric whp, that is

$$\Pr[|\text{Aut}(G_n)| > 1] \xrightarrow{n \to \infty} 0.$$
Structural Entropy for Erdös-Rényi Graphs

**Theorem 2** (Choi, W.S 2009). 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.),

\[ H_S = \binom{n}{2} h(p) - \log n! + O \left( \frac{\log n}{n^a} \right) = \binom{n}{2} h(p) - n \log n + n \log e + O(\log n), \quad a > 1 \]

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

**AEP for structures:** \[ 2^{-\left( \binom{n}{2} (h(p) + \varepsilon) + \log n! \right)} \leq P(S) \leq 2^{-\left( \binom{n}{2} (h(p) - \varepsilon) + \log n! \right)}. \]

**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 $G(n, p)$. 
Outline Update

1. Structural Compression
   - Motivation
   - Unlabeled Graphs
   - SZIP Algorithm and Its Analysis
   - Structural Binary Symmetric Channel

2. Structure of Markov Fields

3. Sequence-Structure Protein Folding Channel
Compression Algorithm called Structural zip, in short SZIP – [Demo].

\[
\begin{align*}
B1 &= 0100110100001110101 \\
B2 &= 1001011000000101
\end{align*}
\]
Asymptotic Optimality of SZIP for Erdős-Rényi Graphs

**Theorem 3** (Choi, W.S., 2012). Let $L(S) = |\tilde{B}_1| + |\tilde{B}_2|$ be the code length.

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

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

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

(iii) The algorithm runs in $O(n + e)$ on average, where $e$ is edges.
Asymptotic Optimality of SZIP for Erdös-Rényi Graphs

Theorem 3 (Choi, W.S., 2012). Let $L(S) = |\tilde{B}_1| + |\tilde{B}_2|$ be the code length.

(i) For large $n$,

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

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

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

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

(iii) The algorithm runs in $O(n + e)$ on average, where $e \# \text{edges}$.

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>Real-world</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<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>1,59,504</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 (Erdös)</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>37,0,18,710</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>
</tbody>
</table>
1. **Structural Compression**
   - Motivation
   - Unlabeled Graphs
   - SZIP Algorithm and Its Analysis
   - Structural Binary Symmetric Channel

2. Structure of Markov Fields

3. Sequence-Structure Protein Folding Channel
Analysis of \textsc{S}z\textsc{ip}: Recurrences for \(\mathbb{E}[B_1]\) and \(\mathbb{E}[B_2]\)

Let \(N_x\) be the number of vertices that passed through node \(x\) in \(T_n\).

\[
\begin{align*}
|B_1| &= \sum_{x \in T_n \text{ and } N_x > 1} \lceil \log(N_x + 1) \rceil \\
|B_2| &= \sum_{x \in T_n \text{ and } N_x = 1} \lceil \log(N_x + 1) \rceil \\
&= \sum_{x \in T_n \text{ and } N_x = 1} 1.
\end{align*}
\]
Analysis of *S*ZIP: Recurrences for $\mathbb{E}[B_1]$ and $\mathbb{E}[B_2]$

Let $N_x$ be the number of vertices that passed through node $x$ in $T_n$.

$$|B_1| = \sum_{x \in T_n \text{ and } N_x > 1} \lceil \log(N_x + 1) \rceil$$

$$|B_2| = \sum_{x \in T_n \text{ and } N_x = 1} \lceil \log(N_x + 1) \rceil$$

$$= \sum_{x \in T_n \text{ and } N_x = 1} 1.$$  

Both $\mathbb{E}[|B_1|]$ and $\mathbb{E}[|B_2|]$ satisfy two-dimensional recurrences for some $d \geq 0$

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

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

for some $c_n$ (e.g., $c_n = \lceil \log(n + 1) \rceil$ or $c_n = n$).
Another Look – \((n, d)\)-tries

1. The root of a tree contains \(n\) balls.
2. Balls independently move down to the left subtree (with probability \(p\)) or the right subtree (with probability \(1 - p\)).
3. For a non-negative integer \(d\), at level \(d\) or greater one ball is removed from the leftmost node.
Another Look – \((n, d)\)-tries

1. The root of a tree contains \(n\) balls.
2. Balls independently move down to the left subtree (with probability \(p\)) or the right subtree (with probability \(1 - p\)).
3. For a non-negative integer \(d\), at level \(d\) or greater one ball is removed from the leftmost node.

For example for \(c_n = n\):

\[
a(n, d) = \frac{1}{h} n \log n + \frac{1}{h} \left[ \gamma + \frac{h_2}{2h} + \phi(\log p, n) \right] n + \frac{1}{2h \log p} \log^2 n + \frac{d}{h} \log n + O(1)
\]

where \(h = -p \log p - q \log q\), \(h_2 = p \log^2 p + q \log^2 q\), \(\gamma\) is the Euler constant, and \(\phi(x)\) is the periodic function.
Outline Update

1. Structural Compression
   - Motivation
   - Unlabeled Graphs
   - ZIP Algorithm and Its Analysis
   - Structural Binary Symmetric Channel

2. Structure of Markov Fields

3. Sequence-Structure Protein Folding Channel
Structural Binary Symmetric Channel (SBSC)
**Example:** Graph $G_1 = \{A, B, C, D\}$ transmitted with output $G_2$.

Adjacency matrices are: $G_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$, $G_2 = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$.

How much **structural information** can be reliably transmitted over a **noisy channel**?
Capacity of SBSC

Capacity of SBSC is defined as

\[ C = \lim_{n \to \infty} \frac{1}{\binom{n}{2}} \max_{0 \leq p \leq 1} I(S; S') \]

where \( I(S; S') \) is the mutual information between the output structure \( S' \) and the input structure \( S \).

**Theorem 4.** Capacity of the structural Binary Symmetric Channel SBSC(\( \epsilon \)) of Erdős-Rényi graphs is

\[ C = 1 - h(\epsilon) \]

where \( \epsilon \) is the error bit rate and

\[ h(\epsilon) = -\epsilon \log \epsilon - (1 - \epsilon) \log(1 - \epsilon) \]

is the binary entropy.
Outline Update

1. Structural Compression

2. **Structure of Markov Fields**
   - One Dimensional Markov Types
   - One-Dimensional Universal Types.
   - Markov Fields and Tilings

3. Sequence-Structure Protein Folding Channel
Large Systems with Local Interactions

These local interactions are often represented by shapes and tiles leading to a Markov field.

**Markov Field Types:**
Two Markov fields have the same type if they have the same empirical distribution.

The method of types is a powerful technique in information theory; it reduces calculations of the probability of rare events to combinatorics.
1. Structural Compression

2. Structure of Markov Fields
   - One Dimensional Markov Types
   - One-Dimensional Universal Types.
   - Markov Fields and Tilings

3. Sequence-Structure Protein Folding Channel
One-Dimensional Markov: Sequences $x^n = x_1 \ldots x_n$ over $\mathcal{A} = \{1, 2, \ldots, m\}$ alphabet. Define $\mathcal{T}_n(x^n) = \{y^n : P(x^n) = P(y^n)\}$, and $\mathcal{P}_n := \mathcal{P}_n(m)$ class of distributions.

Consider a Markov source with the transition matrix $P = \{p_{ij}\}_{i,j=1}^m$. Then

$$P(x_1^n) = p_{11}^{k_{11}} \ldots p_{mm}^{k_{mm}} = \prod_{i,j \in \mathcal{A}} p_{ij}^{k_{ij}} ,$$

where $k_{ij}$ is the number of pair symbols $(ij)$ in $x_1^n$, that is, $i$ followed by $j$. 


Let’s Begin... One-Dimensional Markov Chains

**One-Dimensional Markov:** Sequences $x^n = x_1 \ldots x_n$ over $A = \{1, 2, \ldots, m\}$ alphabet. Define $T_n(x^n) = \{y^n : P(x^n) = P(y^n)\}$, and $\mathcal{P}_n := \mathcal{P}_n(m)$ class of distributions.

Consider a Markov source with the transition matrix $P = \{p_{ij}\}_{i,j=1}^m$. Then

$$P(x^n_1) = p_{11}^{k_{11}} \cdots p_{mm}^{k_{mm}} = \prod_{i,j \in A} p_{ij}^{k_{ij}},$$

where $k_{ij}$ is the number of pair symbols $(ij)$ in $x^n_1$, that is, $i$ followed by $j$.

For circular strings (i.e., after the $n$th symbol we re-visit the first symbol of $x^n_1$), the matrix $k = [k_{ij}]$ satisfies the following constraints denoted as $\mathcal{F}_n(m)$:

$$\sum_{1 \leq i,j \leq m} k_{ij} = n, \quad \sum_{j=1}^m k_{ij} = \sum_{j=1}^m k_{ji}$$

For example: $m=3$

$k_{11} + k_{12} + k_{13} + k_{21} + k_{22} + k_{23} + k_{31} + k_{32} + k_{33} = n$

$k_{12} + k_{13} = k_{21} + k_{31}$

$k_{12} + k_{32} = k_{21} + k_{23}$

$k_{13} + k_{23} = k_{31} + k_{32}$
Example: Let \( A = \{0, 1\} \) and

\[
k = \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}
\]
Example: Let $\mathcal{A} = \{0, 1\}$ and

$$k = \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$$

$\mathcal{P}_n(m)$ – Markov types but also . . .
a set of all connected Eulerian di-graphs $G = (V(G), E(G))$ such that $V(G) \subseteq \mathcal{A}$ and $|E(G)| = n$.

$\mathcal{E}_n(m)$ – set of connected Eulerian digraphs on $\mathcal{A}$.

$\mathcal{F}_n(m)$ – balanced matrices but also . . .
set of (not necessary connected) Eulerian digraphs on $\mathcal{A}$.

Asymptotic equivalence: $|\mathcal{P}_n(m)| = |\mathcal{F}_n(m)| + O(n^{m^2-3m+3}) \sim |\mathcal{E}_n(m)|$. 
Main Results for One-Dimensional Markov Chains

Theorem 5. (i) For fixed $m$ and $n \to \infty$ the number of Markov types is

$$|\mathcal{P}_n(m)| = d(m) \frac{n^{m^2-m}}{(m^2 - m)!} + O(n^{m^2-m-1})$$

where $d(m)$ is a constant that also can be expressed as

$$d(m) = \frac{1}{(2\pi)^{m-1}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{m-1} \frac{1}{1 + \phi_j^2} \prod_{k \neq \ell} \frac{1}{1 + (\phi_k - \phi_\ell)^2} d\phi_1 d\phi_2 \cdots d\phi_{m-1}.$$ 

(ii) When $m \to \infty$ we find that

$$|\mathcal{P}_n(m)| \sim \frac{\sqrt{2} m^{3m/2} e^{m^2}}{m^{2m^2} 2^m \pi^{m/2}} \cdot n^{m^2-m}$$

provided that $m^4 = o(n)$.

Example. The coefficients at $n^{m^2-m}$ are very small. For $m = 4$ the coefficient is $1.767043356 \times 10^{-11}$. 
Outline Update

1. Structural Compression

2. Structure of Markov Fields
   - One Dimensional Markov Types
   - One-Dimensional Universal Types
   - Markov Fields and Tilings

3. Sequence-Structure Protein Folding Channel
Universal Types (Still One-Dimensional)

Seroussi introduced in 2003 universal types for stationary ergodic sources:

\[(0) (1) (00) (10) (11)\]
\[(1) (0) (10) (11) (00)\]

\[(0) (1) (00) (01) (11)\]
\[(1) (0) (01) (11) (00)\]

\[p = \text{path length} = 8\]

Lempel-Ziv’78 parsing scheme of a sequence of length \(p\) can be represented by a binary tree of path length \(p\).

- \(\mathcal{T}_p\) be the set of binary trees with the path length equal to \(p\).

\[\# \text{ universal types over } \mathcal{A}^p \equiv |\mathcal{T}_p|: \# \text{ of trees of a given path length } p.\]

How to enumerate binary trees of a given path length \(p\)?
Let \( b(n, p) \) be the number of binary trees with \( n \) nodes and path length \( p \). It satisfies:

\[
b(n, p) = \sum_{k+\ell=n-1} \sum_{r+s+n-1=p} b(k, r)b(\ell, s)
\]

Define \( B_n(w) = \sum_{p=0}^{\infty} b(n, p)w^p \), and \( B(z, w) = \sum_{n=0}^{\infty} z^n B_n(w) \). Then

\[
B(z, w) = 1 + zB^2(zw, w)
\]

This functional equation is asymmetric with respect to \( z \) and \( w \).
Let \( b(n, p) \) be the number of binary trees with \( n \) nodes and path length \( p \). It satisfies:

\[
\begin{align*}
  b(n, p) &= \sum_{k+\ell=n-1} \sum_{r+s+n-1=p} b(k, r)b(\ell, s)
\end{align*}
\]

Define \( B_n(w) = \sum_{p=0}^{\infty} b(n, p) w^p \), and \( B(z, w) = \sum_{n=0}^{\infty} z^n B_n(w) \). Then

\[
B(z, w) = 1 + zB^2(zw, w)
\]

This functional equation is asymmetric with respect to \( z \) and \( w \).

We want to study the number of trees in \( \mathcal{T}_p \) (of a given path length \( p \)). Observe

\[
|\mathcal{T}_p| = \sum_{n \geq 0} b(n, p) = [w^p]B(1, w).
\]

We set \( z = 1 \) in the functional equation leading to

\[
B(1, w) = 1 + B^2(w, w)
\]

which is not algebraically solvable.
Number of Trees with a Given Path Length

These results are obtained using the WKB method of applied mathematics. Seroussi (2004) and Knessl & W.S (2004) prove that ($c_1$, $c_2$ are constants)

$$|T_p| = \frac{1}{(\log_2 p) \sqrt{\pi p}} 2^{\frac{2p}{\log_2 p}} \left( 1 + c_1 \log^{-2/3} p + c_2 \log^{-1} p + O(\log^{-4/3} p) \right).$$
Number of Trees with a Given Path Length

These results are obtained using the WKB method of applied mathematics.

Seroussi (2004) and Knessl & W.S (2004) prove that \((c_1, c_2 \text{ are constants})\)

\[
|T_p| = \frac{1}{(\log_2 p) \sqrt{\pi p}} 2^{\frac{2p}{\log_2 p}} \left(1 + c_1 \log^{-2/3} p + c_2 \log^{-1} p + O(\log^{-4/3} p)\right).
\]

When randomly selecting a tree from \(T_p\) we may define: \(N_p\), the number of nodes in the \(T_p\)-model. Surprisingly, we can prove that \(N_p\) is asymptotically normal, that is,

\[
\Pr\{N_p = n\} = \frac{b(n, p)}{\sum_{n=0}^{\infty} b(n, p)} \sim \frac{1}{\sqrt{2\pi \Var[N_p]}} \exp\left[-\frac{(n - \E[N_p])^2}{2\Var[N_p]}\right]
\]

where

\[
\E[N_p] \sim \frac{p}{\log_2 p}, \quad \Var[N_p] \sim \frac{p}{\log_2 p^{5/3}} \frac{(\log 2) A_0}{6(2^{1/3})}
\]

where \(A_0\) is a constant.
1. Structural Compression

2. **Structure of Markov Fields**
   - One Dimensional Markov Types
   - **Markov Fields and Tilings**

3. Sequence-Structure Protein Folding Channel
(Cyclic) Markov Fields and Tilings

*d*-Dimensional Markov Fields:
Consider a *d*-dimensional box \((n_1, \ldots, n_d)\) with \(N = n_1 \cdots n_d\).
A circular representation of such a box is a **torus** that we denote as \(\mathcal{O}_n\).
The shape of interaction is \(S \subset \mathbb{Z}^d\).
A tile \(t\) is \(t : S \to A\) and \(T = \{t : S \to A\}\).

**Markov Field Type** \(\mathcal{X}^n = \{x^n : \mathcal{O}_n \to A\} \cap A\):  
Define the frequency vector of dimension \(D = |T| = m^{|S|}\): 
\[
k(t) \equiv k_S(t) = |\{s \in \mathcal{O}_n : x|_{S+s} = t\}|, \quad t \in T.
\]

**Example:**
A set of Markov field types or tile types is:
\[
\mathcal{P}_n(m, S) = \{k : \exists x \in \mathcal{X}_n \ x^n \text{ is of type } k\}.
\]
Conservation Laws:

\[ \forall \emptyset \neq S' \subset S, \ s \in \mathbb{Z}^d: (S' + s) \subset S \quad \forall t': S' \rightarrow A \quad k_{S'}(t') = k_{S' + s}(t') \]

with shift \( s \in \mathbb{Z}^d \) subject to \( (S' + s) \subset S \).

Example:

\[ k \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + k \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + k \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = k \begin{pmatrix} * \\ * \\ * \end{pmatrix} = k \begin{pmatrix} * \\ 0 \\ 0 \end{pmatrix} + k \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + k \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + k \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \]
Conservation Laws

Conservation Laws:

\[ \forall \emptyset \neq S' \subseteq S, \ s \in \mathbb{Z}^d : (S' + s) \subseteq S \ \forall t' : S' \to A \quad k_{S'}(t') = k_{S'+s}(t') \]

with shift \( s \in \mathbb{Z}^d \) subject to \( (S' + s) \subseteq S \).

Example:

The conservation laws can be viewed as linear equations with a \( 1 \times D \) row denoted as \( C((S', s, t')) \).

The matrix \( C^* \) is hugely over determined! Our goal is to find \( C \) such that the conservation laws can be written as

\[ Ck = 0. \]

Example 1. \( d = 1 \)-dimensional Markov over \( A = \{1, 2\} \).

Tiles are \( ((11), (21), (12), (22)) \) and the conservation laws are

\[ k(11) + k(12) = k(1*) = k(*1) = k(11) + k(21), \]
\[ k(21) + k(22) = k(2*) = k(*2) = k(12) + k(22). \]

leading to one conservation law \( k(12) - k(21) = 0 \) that in the matrix form is

\[ (0, -1, 1, 0) \cdot k = 0. \]
Dimension of the Frequency Vectors

The conservation laws make the vector count \( k \in \mathbb{Z}^D \) to reside in a space of dimensionality

\[
\mu = D - \text{rk}(C) - 1.
\]

where \( \text{rk}(C) \) is the rank of matrix \( C \).

**Theorem 6.** The matrix \( C \) has rank

\[
\text{rk}(C) = \sum_{S' \in S^0} (|\{s : (S' + s) \subset S\}| - 1)(m - 1)^{|S'|}
\]

and consists of a complete set of linearly independent rows.

In particular, for the box shape \( S = I_{l_1} \times I_{l_2} \times \ldots \times I_{l_d} \) we find

\[
\mu = D - 1 - \text{rk}(C) = \sum_{s \in \{0,1\}^d} m^{\prod_i (l_i - s_i)} \cdot (-1)^{\sum_i s_i}.
\]
Dimension of the Frequency Vectors

The conservation laws make the vector count \( k \in \mathbb{Z}^D \) to reside in a space of dimensionality

\[
\mu = D - \text{rk}(C) - 1.
\]

where \( \text{rk}(C) \) is the rank of matrix \( C \).

**Theorem 6.** The matrix \( C \) has rank

\[
\text{rk}(C) = \sum_{S' \in S^0} (|\{s : (S' + s) \subset S\}| - 1)(m - 1)^{|S'|}
\]

and consists of a complete set of linearly independent rows.

In particular, for the box shape \( S = I_{l_1} \times I_{l_2} \times \ldots \times I_{l_d} \) we find

\[
\mu = D - 1 - \text{rk}(C) = \sum_{s \in \{0,1\}^d} m^{\prod_i (l_i - s_i)} \cdot (-1)^{\sum_i s_i}.
\]

**Example:** For \( d = 2 \) and a \( 2 \times 2 \) square shape we have \( \mu = m^4 - 2m^2 + m \), while for a \( 3 \times 2 \) rectangular shape we find \( \mu = m^6 - m^4 - m^3 + m^2 \).
We view the count vector $\mathbf{k} = \{k(t)\}_{t \in T}$ in the $D = m^{|S|}$ space.

\[
\mathcal{C} = \{\mathbf{k} \in \mathbb{N}^D : C_m(S) \cdot \mathbf{k} = 0\}
\]
\[
\mathcal{F} = \{\mathbf{k} \in \mathcal{C} : \sum_i k^i = N\}
\]
\[
\hat{\mathcal{F}} = \{\hat{\mathbf{k}} \in \{(\mathbb{R}^D_{+} : C \cdot \hat{\mathbf{k}} = 0, \sum_i \hat{k}_i = 1\}
\]
\[
\mathcal{F}_N = \{N\hat{\mathbf{k}} : \hat{\mathbf{k}} \in \hat{\mathcal{F}}, N\hat{\mathbf{k}} \in \mathbb{Z}^D\}
\]

The polytope $\mathcal{F}_N$ is of dimension $\mu$.

**Topological Closure** of (normalized) $\hat{\mathcal{P}}$ is a convex subset of $\hat{\mathcal{F}}$. 
We view the count vector $k = \{k(t)\}_{t \in T}$ in the $D = m^{|S|}$ space.

The polytope $\mathcal{F}_N$ is of dimension $\mu$.

**Topological Closure** of (normalized) $\hat{P}$ is a convex subset of $\hat{F}$.

The lattice $\mathcal{F}_N$ consists of all integer points inside $\hat{F}$ scaled by $N$. Volume of $\mathcal{F}_N$ is of order $N^\mu$ with integer points growing as $N^\mu$.

**Theorem 7** (Ehrhart, 1967). If $\hat{F}$ is a convex polytope with vertices in $\mathbb{Q}^D$, where $\mathbb{Q}$ is the set of rational numbers, then that $c_{\mu,j} \neq 0$ for some $j$

\[
|\mathcal{F}_N| = a_{\mu,j} N^\mu + a_{\mu-1,j} N^{\mu-1} + \ldots + a_{0,j} N \equiv j \pmod{p}.
\]
Main Result for Markov Types

**Theorem 8.** Consider the torus $\mathcal{O}_n$. There exists $0 < c_{\text{min}} \leq c_{\text{max}}$ such that

$$c_{\text{min}} N^\mu \leq |\mathcal{P}_n(m, S)| \leq c_{\text{max}} N^\mu$$
Main Result for Markov Types

**Theorem 8.** Consider the torus $O_n$. There exists $0 < c_{min} \leq c_{max}$ such that

$$c_{min} N^\mu \leq |\mathcal{P}_n(m, S)| \leq c_{max} N^\mu$$

**Lemma 3.** There exist $\mu + 1$ linearly independent periodic tilings.

![Diagram](image_url)

- take all possible markings inside:
- there would be $m^{|S|}$ of them,
- but some appear a few times:

$m = 2$:

1. 2

these appear more than once:

- 6
- 4
- 3
- 2 appearances

$2^6 - 5 - 3 - 2 - 7 \cdot 1 = 64 - 19 = 45$ essentially different
Main Result for Markov Types

**Theorem 8.** Consider the torus $\mathcal{O}_n$. There exists $0 < c_{\min} \leq c_{\max}$ such that

$$c_{\min} N^\mu \leq |\mathcal{P}_n(m, S)| \leq c_{\max} N^\mu$$

**Lemma 3.** There exist $\mu + 1$ linearly independent periodic tilings.

---

Take all possible markings inside:
- there would be $m^{|S|}$ of them,
- but some appear a few times:

For $m = 2$:
- these appear more than once:

\[
\begin{align*}
2^6 - 5 - 3 - 2 - 7 \cdot 1 &= 64 - 19 = 45 \quad \text{essentially different}
\end{align*}
\]

---

Each of shape is one of $\mu + 1$ size base

\[
\begin{align*}
\frac{a_1 k_1 + \ldots + a_{\mu+1} k_{\mu+1}}{a_1 + \ldots + a_{\mu+1}} : \forall_i a_i \in \mathbb{N}, \sum_i a_i = N \\
\binom{N + \mu - 1}{\mu} = O(N^\mu)
\end{align*}
\]
Outline Update

1. Structural Compression
2. Structure of Markov Fields
3. Sequence-Structure Protein Folding Channel
**Sequence-Structure Channel**

\[ p(f|s) := \frac{e^{-\beta \mathcal{E}(s,f)}}{Z(s, \beta)} \quad \text{and} \quad Z(s, \beta) := \sum_{f \in \mathcal{F}} e^{-\beta \mathcal{E}(s,f)} \]
Sequence-Structure Channel

\[ p(f|s) := \frac{e^{-\beta \mathcal{E}(s,f)}}{Z(s, \beta)} \]

\[ Z(s, \beta) := \sum_{f \in \mathcal{F}} e^{-\beta \mathcal{E}(s,f)} \]

**Capacity:**

\[ C = \max_{p(s)} I(S; F) \sim \log |F_N| - \min_{p(s)} H(F|S). \]

where \( F_N \): set of self-avoiding walks of length \( N \).

**Conditional Entropy**

\[ H(S|F) = \mathbb{E}[\log Z(S, \beta)] + \beta \mathbb{E}[\mathcal{E}(F, S)] \]

where \( \mathcal{E}(F, S) \) is energy of a walk \( F \) over sequence \( s \).

Clearly, \( \mathbb{E}[\mathcal{E}(F, S)] = N \cdot \alpha(S) \) for some \( \alpha(S) \).
Motivation & Experimental Results

Protein Folds in Nature

Probability of protein folds vs. sequence rank.
Motivation & Experimental Results

Protein Folds in Nature

Probability of protein folds vs. sequence rank.

Optimal input distribution from Blahut-Arimoto algorithm:
Phase Transition

Define free energy $\gamma(\beta, S)$ as

$$
\gamma(\beta, S) = \lim_{N \to \infty} \frac{\mathbf{E}[\log Z(S, \beta)]}{\log |\mathcal{F}_N|}.
$$

By sub-multiplicative property of $F_N$ we can prove there exists $\mu$ such that

$$
\frac{\log |F_N|}{N} \xrightarrow{N \to \infty} \log \mu.
$$

Then

$$
\mathbf{E} \log Z(S, \beta) \sim \log |F_N| \cdot \gamma(\beta, S) \sim N \log \mu \cdot \gamma(\beta, S)
$$

leading to

$$
H(F|S) \sim N[\gamma(\beta, S) \log \mu + \beta \alpha(S)]
$$

Phase transition of the free energy $\gamma(\beta, S)$ (hence capacity $C$) leads to

$$
\log \mu \cdot \gamma(\beta, S) = \begin{cases} 
\log \mu - \beta \alpha + \frac{1}{2} \sigma^2 \beta^2 & \beta < \frac{\sqrt{2 \log \mu}}{\sigma} \\
\beta \sqrt{2 \sigma^2 \log \mu} - \beta \alpha & \beta \geq \frac{\sqrt{2 \log \mu}}{\sigma}
\end{cases}
$$
Experimental Confirmation of the Phase Transition
That’s It

THANK YOU