Structural Information

Wojciech Szpankowski
Purdue University
W. Lafayette, IN 47907

September 26, 2016

ETH, Zurich, 2016

Structural Information

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.

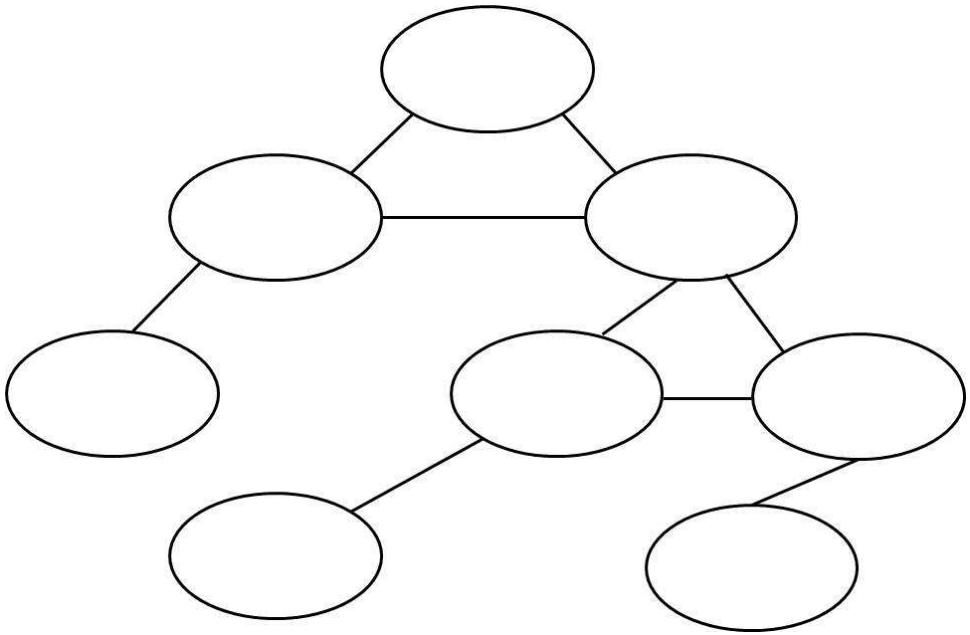
Outline

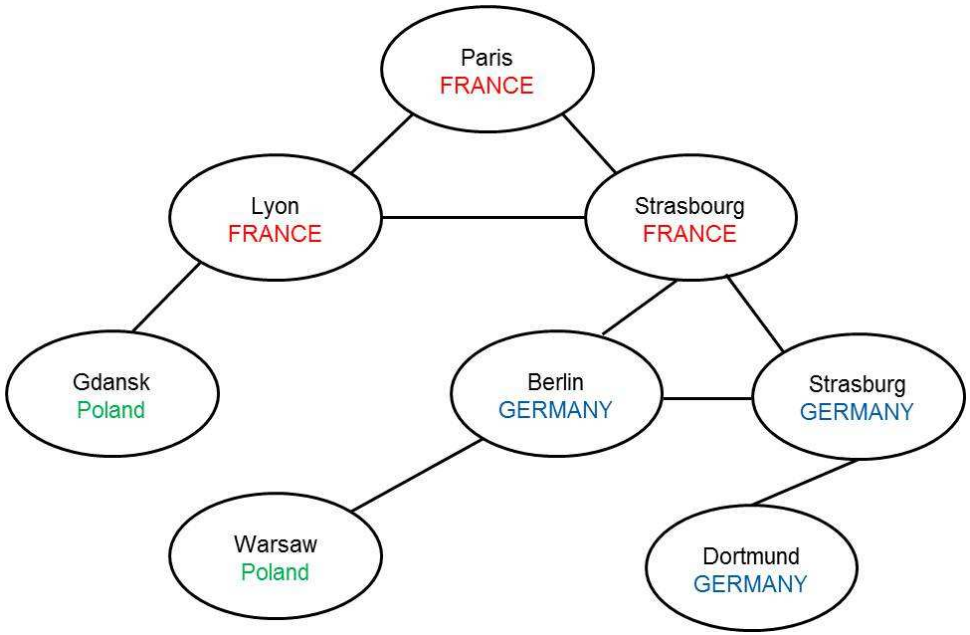
1. Structural Compression
   - Motivation
   - Unlabeled Graphs
   - SZIP Algorithm and Its Analysis
   - Structural Binary Symmetric Channel
2. Tree Compression with Correlated Vertex Names
   - Motivation
   - Plane vs Non-Plane Trees
   - Entropy Computation
   - Lossless Compression Algorithms
3. Boltzmann Sequence-Structure Channel
   - Boltzmann Channel
   - Motivation
   - Phase Transition

Outline

1. Structural Compression
   - Motivation
   - Unlabeled Graphs
   - SZIP Algorithm and Its Analysis
   - Structural Binary Symmetric Channel
2. Tree Compression with Correlated Vertex Names
   - Motivation
   - Plane vs Non-Plane Trees
   - Entropy Computation
   - Lossless Compression Algorithms
3. Boltzmann Sequence-Structure Channel
   - Boltzmann Channel
   - Motivation
   - Phase Transition
4. Structure of Markov Fields
   - Markov Types
   - One-Dimensional Markov Chains
   - Markov Fields and Tilings

Graphs with Locally Correlated Labels





How many bits are required to describe the unlabeled graph on the left, and how many additional bits one needs to represent the correlated labels on the right?

The Real Stuff ...

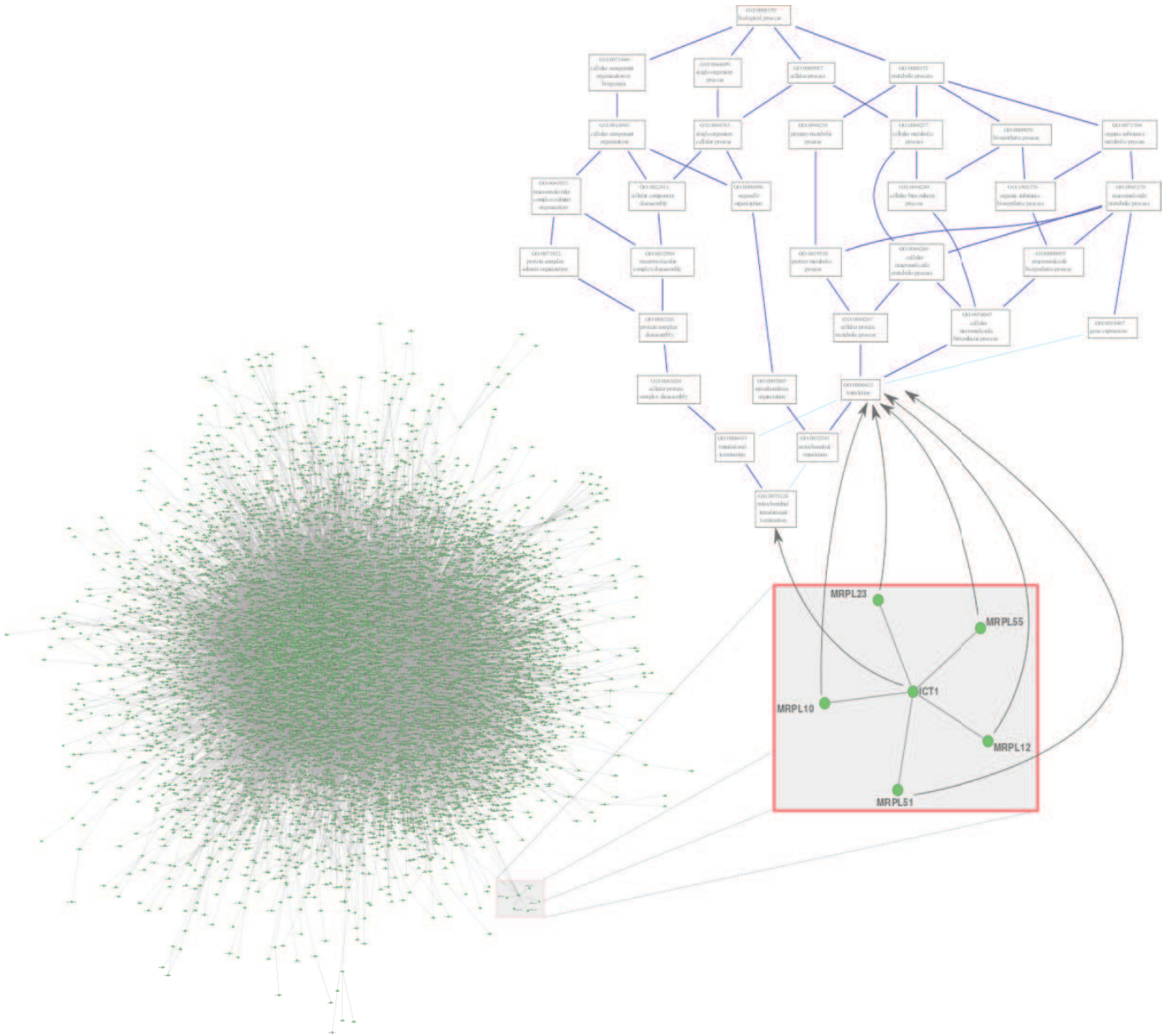


Figure 1: Protein-Protein Interaction Network with BioGRID database

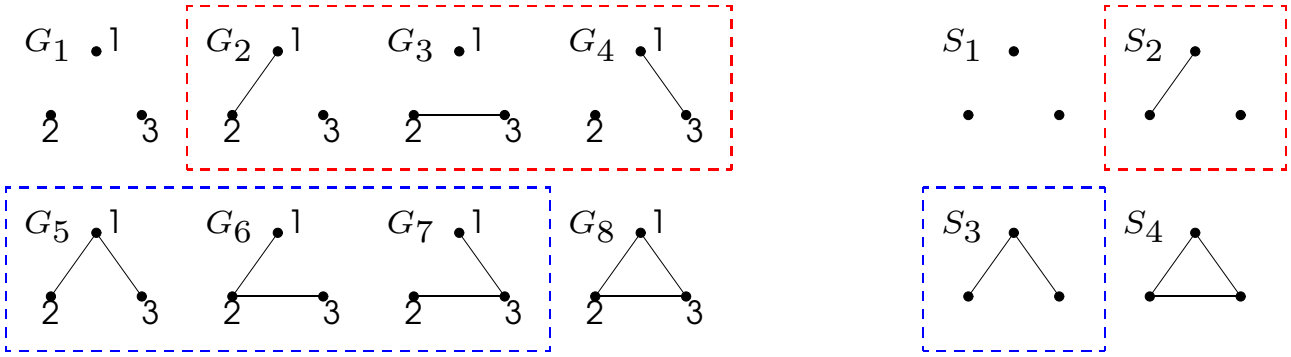
Outline Update

1. Structural Compression
   - Motivation
   - Unlabeled Graphs
   - SZIP Algorithm and Its Analysis
   - Structural Binary Symmetric Channel
2. Tree Compression
3. Sequence-Structure Protein Folding Channel

Graph and Structural Entropies

Information Content of Unlabeled Graphs:

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



Graph Entropy vs **Structural Entropy**:

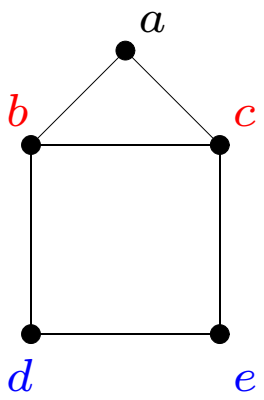
The probability of a structure S is: $P(S) = N(S) \cdot P(G)$
where $N(S)$ is the number of different labeled graphs having the same structure.

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

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

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

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 (i.e., graph perspective is the same).

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 and Symmetry

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

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

If a graph G in $\mathcal{G}(n, p)$ 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 Erdös-Rényi graphs and all p satisfying*

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

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

$$P(\mathrm{Aut}(G) = 1) \sim 1 - O(n^{-w}).$$

Symmetry of Preferential Attachment Graphs?

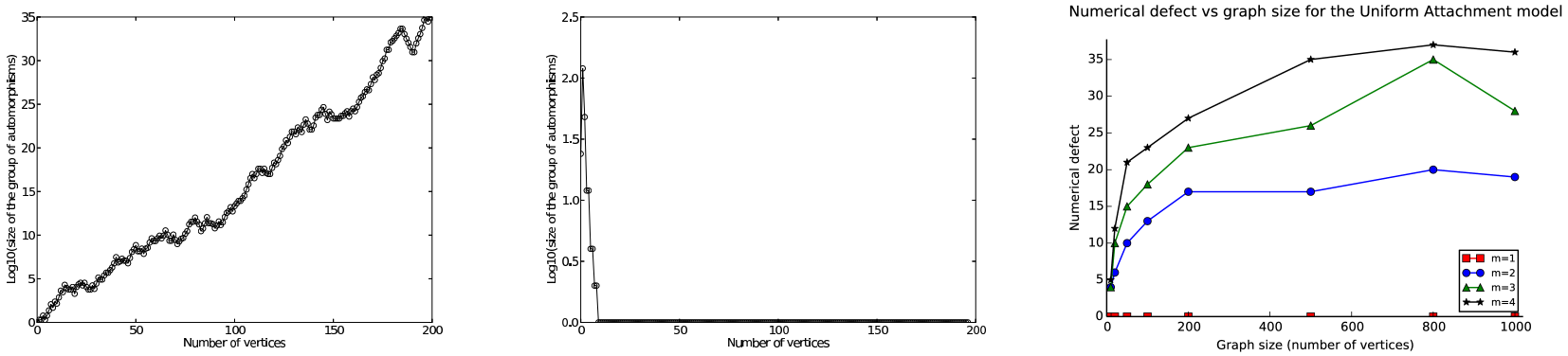
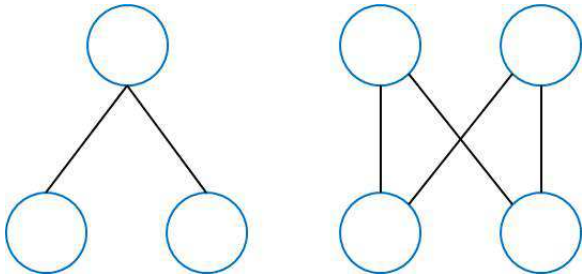


Figure 2: $|\mathrm{Aut}(G)|$: For $m = 1$ (left), $m = 4$ (middle), defect (right).

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*

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

for some $C > 0$.



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

$$\Pr[|\mathrm{Aut}(G_n)| > 1] \xrightarrow{n\to\infty} 0.$$

Symmetry of Preferential Attachment Graphs?

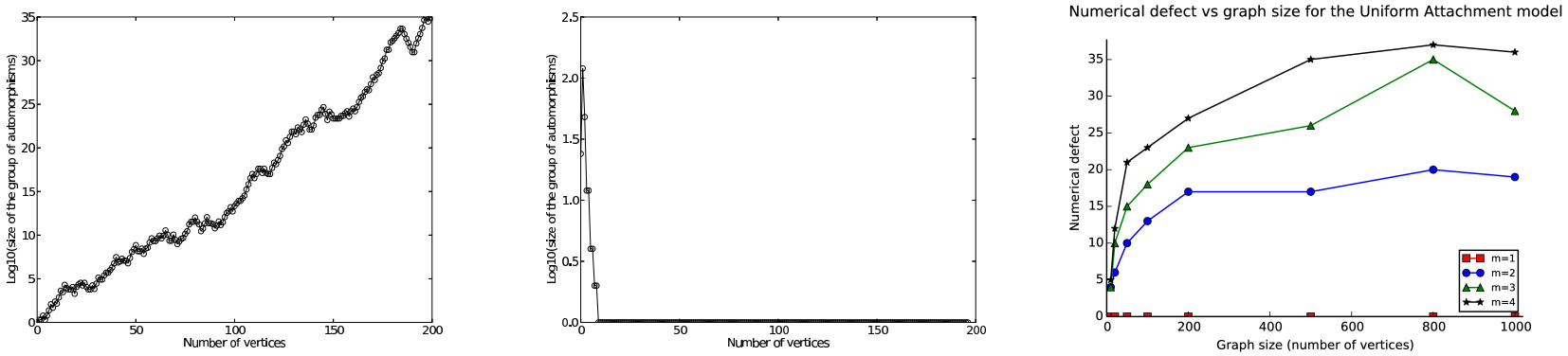
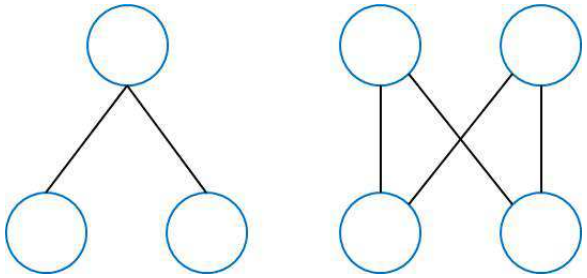


Figure 3: $|\mathrm{Aut}(G)|$: For $m = 1$ (left), $m = 4$ (middle), defect (right).

Theorem 2 (Symmetry Results for $m = 1, 2$). *Let graph G_n be generated by the preferential model with parameter $m = 1$ or $m = 2$. Then*

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

for some $C > 0$.



Theorem (Luczak, Magner, WS, 2016) *For $m \geq 3$ a graph G_n generated by the preferential model is asymmetric whp, that is*

$$\Pr[|\mathrm{Aut}(G_n)| > 1] \xrightarrow{n\to\infty} 0.$$

Structural Entropy for Erdös-Rényi Graphs

Theorem 3 (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.), for some $a > 0$*

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

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:

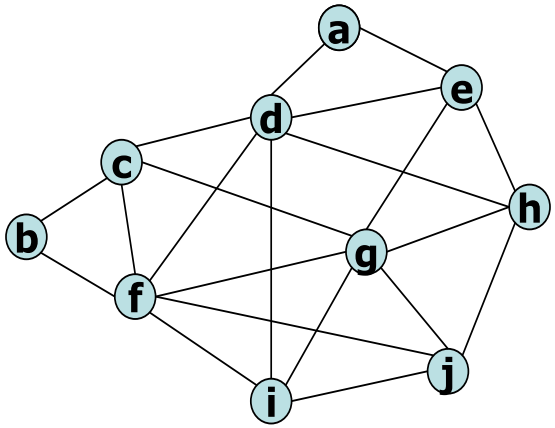
1. $H_{\mathcal{S}} = H_{\mathcal{G}} - \log n! + \sum_{S \in \mathcal{S}} P(S) \log |\mathrm{Aut}(S)|$.
2. $H_{\mathcal{G}} = \binom{n}{2} h(p)$
3. $\sum_{S \in \mathcal{S}} P(S) \log |\mathrm{Aut}(S)| = o(1)$ by asymmetry of $\mathcal{G}(n,p)$.

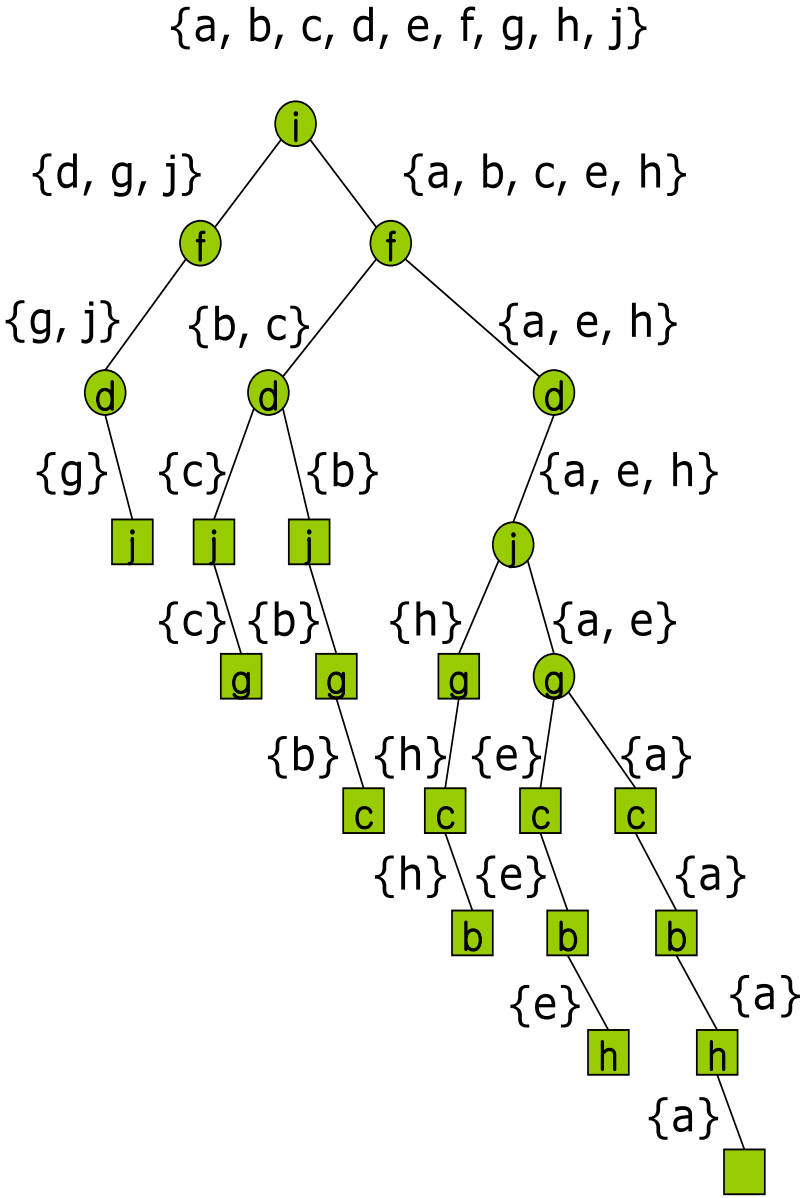
Outline Update

1. Structural Compression
   - Motivation
   - Unlabeled Graphs
   - SZIP Algorithm and Its Analysis
   - Structural Binary Symmetric Channel
2. Tree Compression
3. Sequence-Structure Protein Folding Channel

Structural Zip (SZIP) Algorithm

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





B1 = 0100110100001110101

B2 = 1001011000000101

Asymptotic Optimality of SZIP for Erdös-Rényi Graphs

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

(i) *For large* n,

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

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

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

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

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

Asymptotic Optimality of SZIP for Erdös-Rényi Graphs

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

(i) *For large* n,

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

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

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

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

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

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

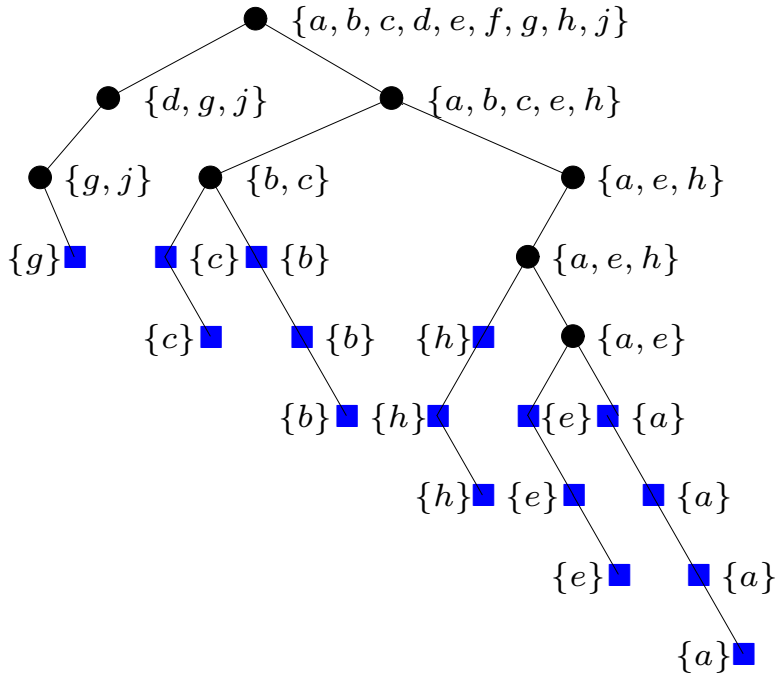
Analysis of SZIP: Recurrences for $\mathbf{E}[B_1]$ and $\mathbf{E}[B_2]$

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

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

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

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



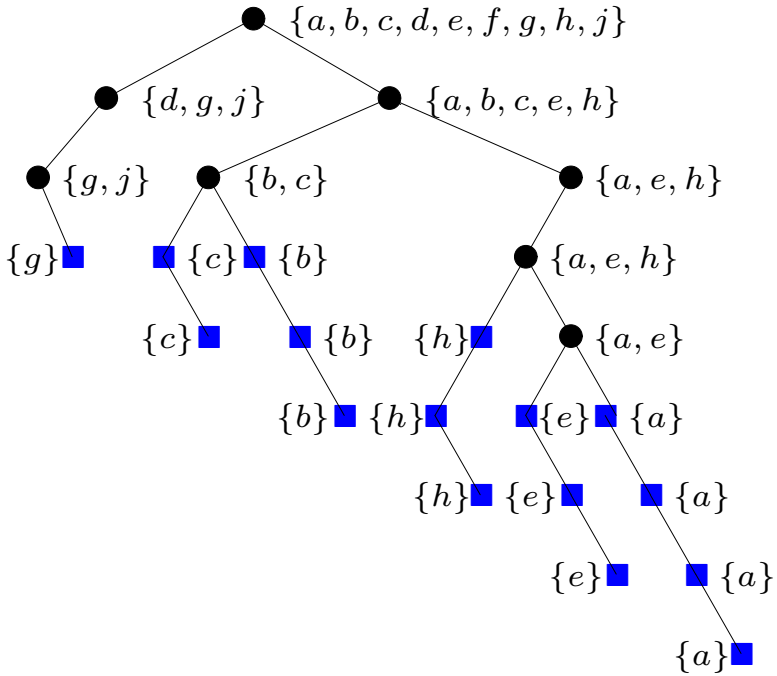
Analysis of SZIP: Recurrences for $\mathbf{E}[B_1]$ and $\mathbf{E}[B_2]$

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

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

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

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

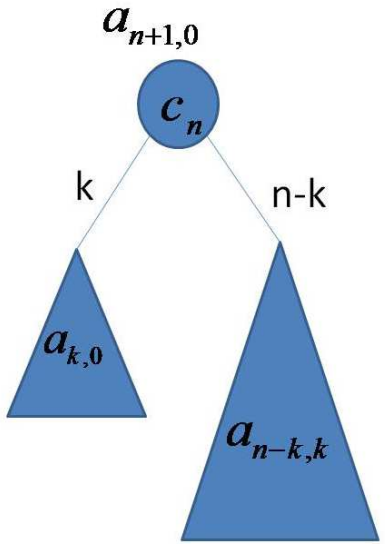


Both $\mathbf{E}[|B_1|]$ and $\mathbf{E}[|B_2|]$ satisfy two-dimensional recurrences for some $d \geq 0$:

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

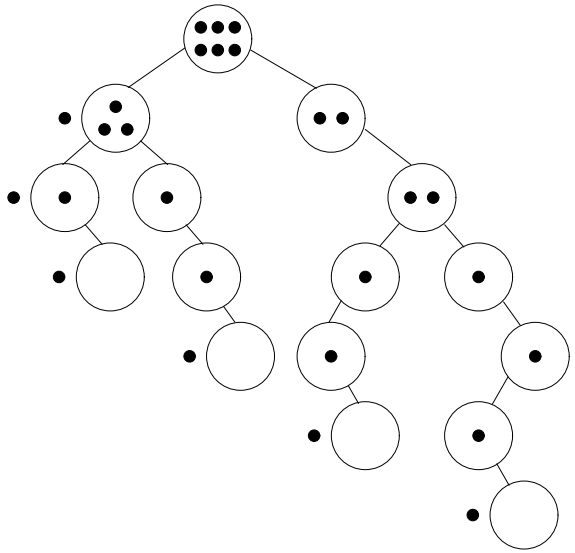
$$a_{n,d} = c_n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (a_{k,d-1} + a_{n-k,k+d-1}).$$

for some c_n (e.g., $c_n = \lceil \log(n+1) \rceil$ or $c_n = n$).



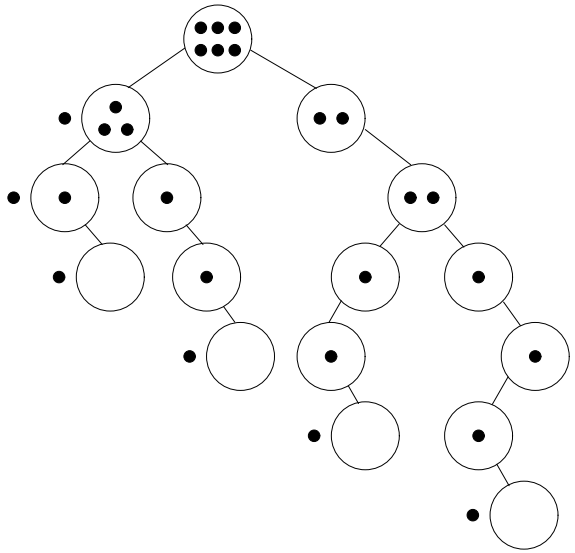
Another Look – (n, d)-tries

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 $c_n = n$:

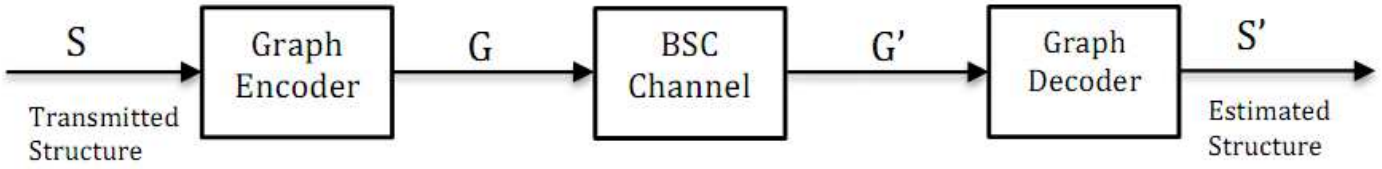
$$a(n,d) = \frac{1}{h} n \log n + \frac{1}{h}\left[\gamma + \frac{h_2}{2h} + \Phi(\log_p n)\right] n + \frac{1}{2h \log p} \log^2 n + \frac{d}{h} \log n + O(1)$$

where $h = -p \log p - q \log q$, $h_2 = p \log^2 p + q \log^2 q$, γ is the Euler constant, and $\Phi(x)$ is the periodic function.

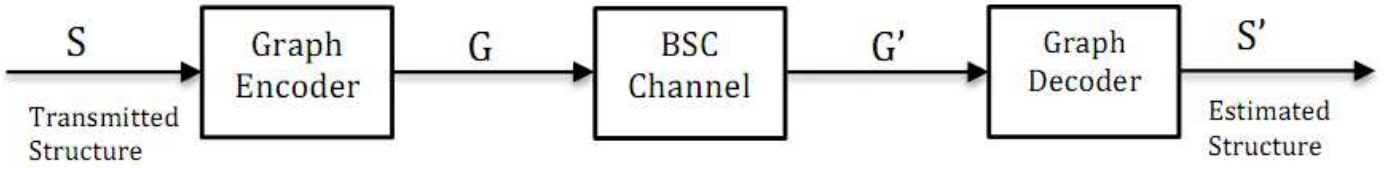
Outline Update

1. Structural Compression
   - Motivation
   - Unlabeled Graphs
   - SZIP Algorithm and Its Analysis
   - Structural Binary Symmetric Channel
2. Tree Compression
3. Sequence-Structure Protein Folding Channel

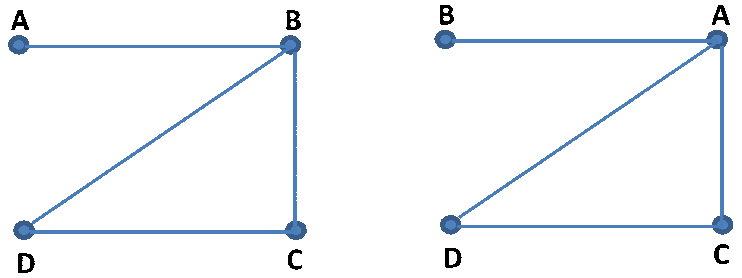
Structural Binary Symmetric Channel (SBSC)



Structural Binary Symmetric Channel (SBSC)



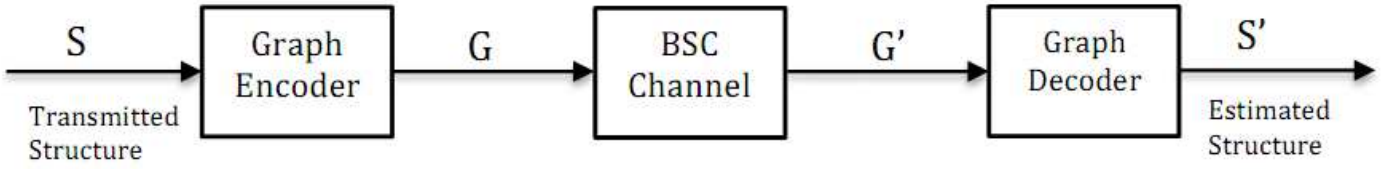
Example: Graph $G_1 = \{A, B, C, D\}$ transmitted with output G_2.



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

How much structural information can be reliably transmitted over a noisy channel?

Capacity of SBSC



Capacity of SBSC is defined as

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

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

Theorem 5. *Capacity of the the structural Binary Symmetric Channel SBSC(ϵ) of Erdös-Rényi graphs is*

$$C = 1 - h(\epsilon)$$

where ε is the error bit rate and

$$h(\varepsilon) = -\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon)$$

is the binary entropy.

Outline

1. Structural Compression
2. Tree Compression with Correlated Vertex Names
   - Motivation
   - Plane vs Non-Plane Trees
   - Entropy Computation
   - Lossless Compression Algorithms
3. Boltzmann Sequence-Structure Channel

Tree with Correlated Names: Motivating Example

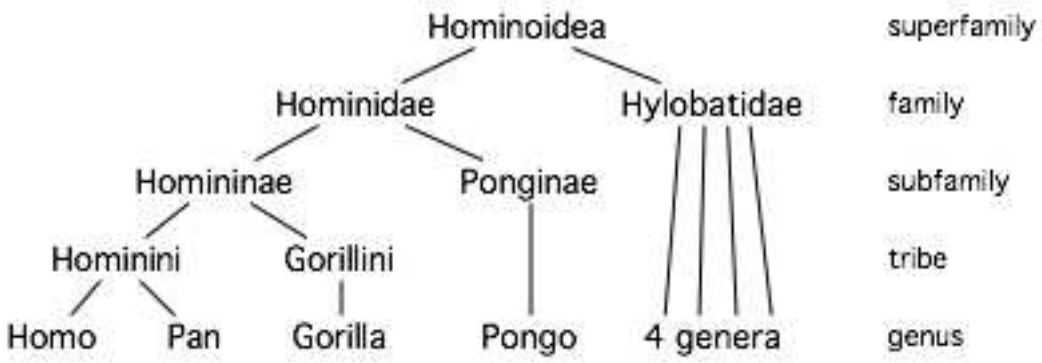


Figure 4: Linnaean taxonomy of *hominoid*

Data in modern applications has various types of structure:

- Topographic maps
- Social/biological networks
- Phylogenetic trees

Tree with Correlated Names: Motivating Example

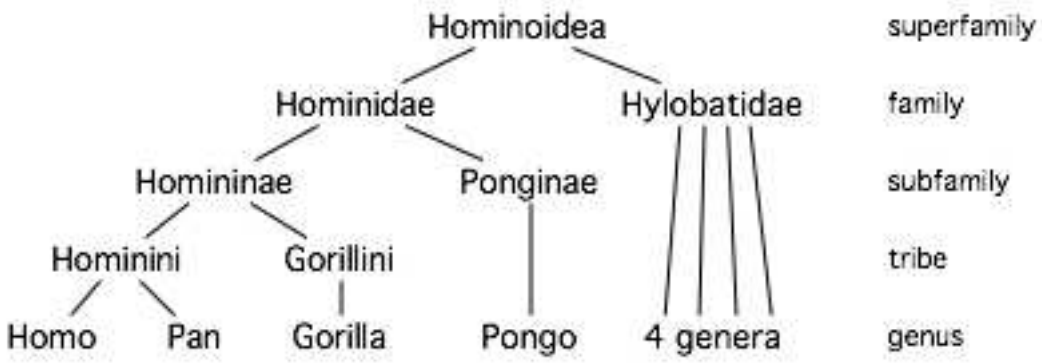


Figure 4: Linnaean taxonomy of *hominoid*

Data in modern applications has various types of structure:

- Topographic maps
- Social/biological networks
- Phylogenetic trees

This motivates the study of information content/compression of data structures with correlated names or in general multimodal data structures.

Source models for trees

Probabilistic models for rooted binary plane trees:

Random binary trees on n leaves:

- At time $t = 0$: Add a node.
- At time $t = 1, \ldots, n$: Choose a leaf uniformly at random and attach 2 children.

Source models for trees

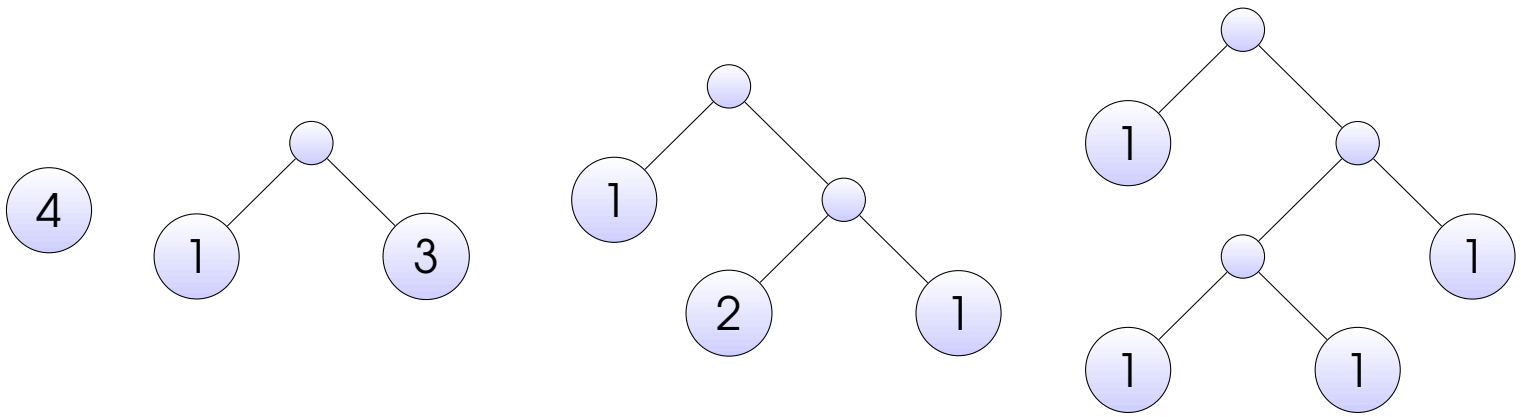
Probabilistic models for rooted binary plane trees:

Random binary trees on n leaves:

- At time $t = 0$: Add a node.
- At time $t = 1, \ldots, n$: Choose a leaf uniformly at random and attach 2 children.

Equivalent formulation:

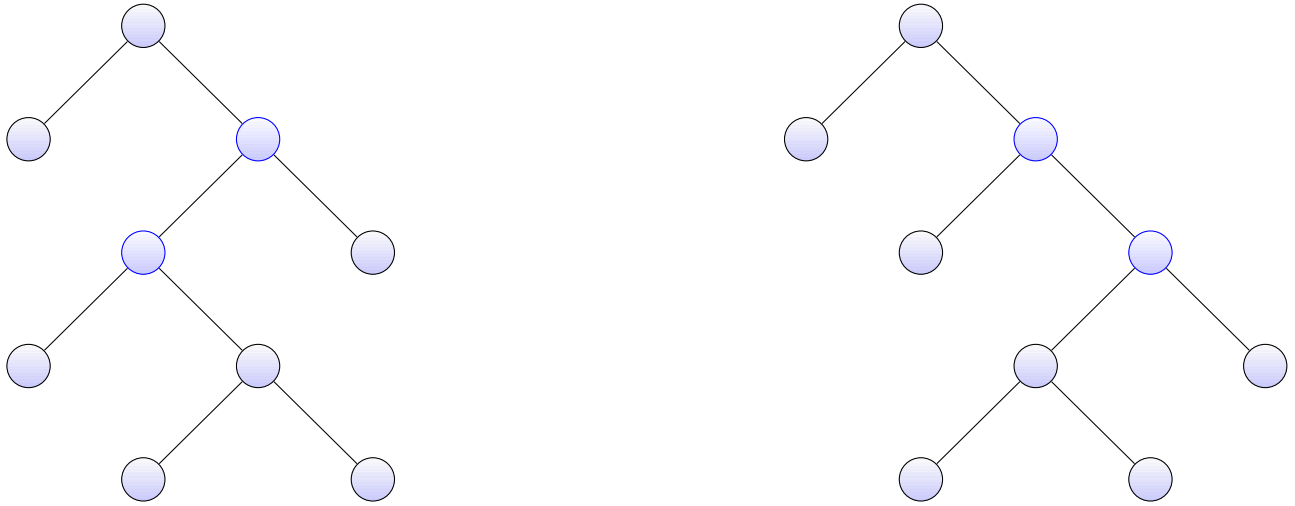
- Initially, add a node with label n.
- While there is a leaf with label $\ell > 1$, choose a number ℓ' uniformly at random from $[\ell - 1]$ and add a left and right child with labels with ℓ' and $\ell - \ell'$, respectively.



Source Models for Non-Plane Trees

Non-plane trees: Ordering of siblings doesn't matter. Formally, a non-plane tree is an equivalence class of trees, where two trees are equivalent if one can be converted to the other by a sequence of rotations.

Example of two equivalent trees:



Source models for vertex names

Parameters for vertex names:

- $\mathcal{A}$: The (finite) alphabet.
- $m \geq 0$: the length of a name.
- P: transition matrix for a Markov chain with state space $\mathcal{A}$.
- π: stationary distribution associated with P.

Generating vertex names given a tree structure:

- Generate a name for the root by taking m letters from a memoryless source with distribution π on $\mathcal{A}$.
- Given a name $a_1 a_2, \ldots, a_m$ for an internal node, generate names for its two children $b_1, \ldots b_m$ and $b'_1, \ldots, b'_m$ such that the jth letter, $j = 1, \ldots, m$, of each child, is generated according to the distribution $P(b_j | a_j)$.

LT_n: a binary plane tree on n leaves with vertex names generated as above.

Entropy for Plane-Oriented Trees with Names

Theorem 6. *The entropy of a plane tree with names, generated according to the model with fixed length m, is given by*

$$H(LT_n) = \log_2(n-1) + 2n \sum_{k=2}^{n-1} \frac{\log_2(k-1)}{k(k+1)} + 2(n-1)mh(P) + mh(\pi)$$

$$= n \cdot \left(2 \sum_{k=2}^{n-1} \frac{\log_2(k-1)}{k(k+1)} + 2mh(P) \right) + O(\log n).$$

where $h(\pi) = -\sum_{a \in \mathcal{A}} \pi(a) \log \pi(a)$.

- $\log_2(n-1)$: The choice of the number of leaves in the left subtree of the root.
- $2n \sum_{k=2}^{n-1} \frac{\log_2(k-1)}{k(k+1)}$: The accumulated choices of the number of leaves in left subtrees.
- $2(n-1)mh(P)$: The choices of vertex names given those of their parents.
- $mh(\pi)$: The choice of the vertex name for the root.

Sketch of Proof

Observe that

$$H(LT_n|F_n(r)) = \log_2(n-1) + 2mh(P) + \frac{2}{n-1}\sum_{k=1}^{n-1} H(LT_k|F_k(r))$$

and $H(LT_n) = H(LT_n|F_n(r)) + H(F_n(r))$, where $F_n(r)$ is the name assigned to the root r.

Sketch of Proof

Observe that

$$H(LT_n|F_n(r)) = \log_2(n-1) + 2mh(P) + \frac{2}{n-1}\sum_{k=1}^{n-1} H(LT_k|F_k(r))$$

and $H(LT_n) = H(LT_n|F_n(r)) + H(F_n(r))$, where $F_n(r)$ is the name assigned to the root r.

The above recurrence has a simple solution as shown in the lemma below.

Lemma 3. *The recurrence* $x_1 = 0$,

$$x_n = a_n + \frac{2}{n-1}\sum_{k=1}^{n-1} x_k, \quad n \geq 2$$

has the following solution for $n \geq 2$*:*

$$x_n = a_n + n\sum_{k=2}^{n-1} \frac{2a_k}{k(k+1)}.$$

Entropy for Non-plane Trees

Entropy for non-plane trees is more difficult: let S_n denote a random non-plane tree on n leaves according to our model.

Theorem 7 (Entropy rate for non-plane trees)**.**

$$H(S_n) = (h(t) - h(t|s)) \cdot n + o(n) \approx 1.109n$$

where

$$h(t) = 2\sum_{k=1}^{\infty} \frac{\log_2 k}{(k+1)(k+2)}, \qquad h(t|s) = 1 - \sum_{k=1}^{\infty} \frac{b_k}{(2k-1)k(2k+1)},$$

and (the coincidence probability)

$$b_k = \sum_{t_k \in \mathcal{T}_k} (\Pr[T_k = t_k])^2.$$

Remark: It turns out that b_n satisfies for $n \geq 2$ the following recurrence

$$b_n = \frac{1}{(n-1)^2} \sum_{j=1}^{n-1} b_j b_{n-j}$$

with $b_1 = 1$.

Remark. The sequence b_k is related to the Rényi entropy of order 1 of T_k.

Sketch of Proof

1. Observe that $H(T_n) - H(S_n) = H(T_n|S_n)$.

Sketch of Proof

1. Observe that $H(T_n) - H(S_n) = H(T_n|S_n)$.

2. For $s \in \mathcal{S}$ and $t \in \mathcal{T}$: $t \sim s$ means the plane tree t is isomorphic to s.
We write: $[s] = \{t \in \mathcal{T} : t \sim s\}$.
3. We have

$$\Pr(S_n = s) = |[s]| \Pr(T_n = t), \quad \Pr(T_n = t|S_n = s) = 1/|[s]|.$$

4. $X(t)$: number of internal vertices of t with unbalanced subtrees;
$Y(t)$: number of internal vertices with balanced, non isomorphic subtrees.
Since $|[s]| = 2^{X(s)+Y(s)}$, thus

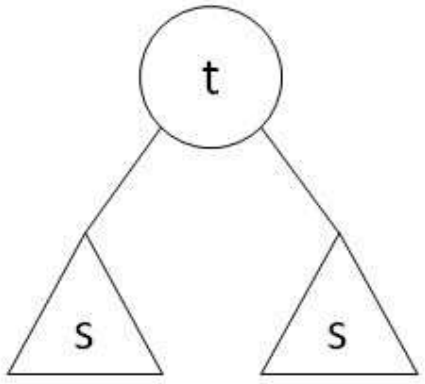
$$H(T_n|S_n) = -\sum_{t\in\mathcal{T}_n, s\in\mathcal{S}_n} \Pr(T_n = t, S_n = s) \log \Pr(T_n = t|S_n = s) = \mathbf{E}X_n + \mathbf{E}Y_n$$

5. Let $Z(t)$ be number of internal vertices of t with isomorphic subtrees. Obviously, $X(t) + Y(t) + Z(t) = n - 1$. Let $Z_n(t) = \sum_{\mathfrak{s}} Z_n(\mathfrak{s})$. Then

$$\mathbf{E}Z_n(\mathfrak{s}) = \mathbf{E}I\left(T_n \sim \mathfrak{s} * \mathfrak{s}\right) + \frac{2}{n-1}\sum_{k=1}^{n-1} \mathbf{E}Z_k(\mathfrak{s})$$

where

$$\mathbf{E}I\left(T_n \sim \mathfrak{s} * \mathfrak{s}\right) = I\left(n = 2\Delta(\mathfrak{s})\right) \frac{\Pr^2(T_{n/2} \sim \mathfrak{s})}{n-1}.$$



Compression algorithms: Arithmetic Encoding

Optimal algorithm for compression of LT_n: arithmetic encoding.

$$\Pr[LT_n = lt] = \Pr[\text{ name of root}] \cdot \Pr[\#\text{ of nodes in left subtree}]$$
$$\cdot \Pr[LT_n^L = lt^L | \text{root name}, \#\text{ of nodes in left subtree}]$$
$$\cdot \Pr[LT_n^R = lt^R | \text{root name}, \#\text{ of nodes in left subtree}]$$

Expected code length: $\mathbf{E}[C_n^{(1)}] \leq H(LT_n) + 2$.

Compression algorithms: Arithmetic Encoding

Optimal algorithm for compression of LT_n: arithmetic encoding.

$$\Pr[LT_n = lt] = \Pr[\text{ name of root}] \cdot \Pr[\#\text{ of nodes in left subtree}]$$
$$\cdot \Pr[LT_n^L = lt^L | \text{root name, \# of nodes in left subtree}]$$
$$\cdot \Pr[LT_n^R = lt^R | \text{root name, \# of nodes in left subtree}]$$

Expected code length: $\mathbf{E}[C_n^{(1)}] \leq H(LT_n) + 2$.

Suboptimal algorithm for non-plane tree compression: Naïve tweak of the algorithm for LT_n. At each stage, encode the smaller subtree first.

Expected code length: $\mathbf{E}[C_n^{(2)}] \leq H(S_n) + 2 + \mathbf{E}[Y_n]$.

Y_n: the number of internal nodes with balanced but not isomorphic subtrees. Exact (but complicated) series formula for $\mathbf{E}[Y_n]$ yields

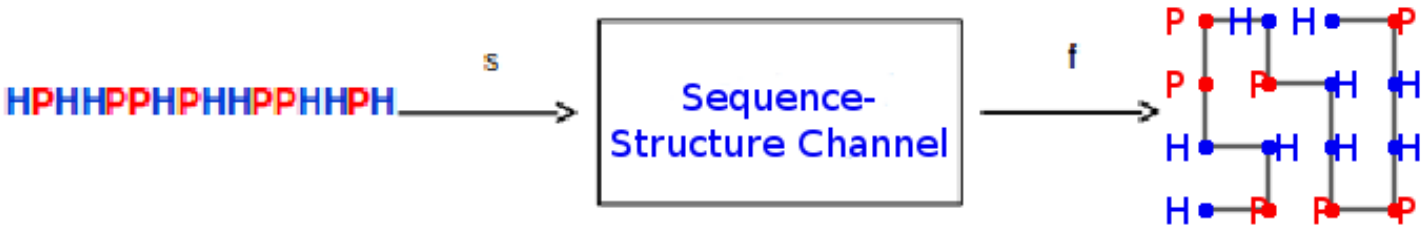
$$\mathbf{E}[C_n^{(2)}] \leq 1.013 H(S_n).$$

Optimal algorithm for non-plane tree compression: $\mathbf{E}[C_n^{(3)}] \leq H(S_n) + 2$. Better compression, worse running time $O(n^2)$.

Outline Update

1. Structural Compression
2. Tree Compression with Correlated Vertex Names
3. Boltzmann Sequence-Structure Channel
   - Boltzmann Channel
   - Motivation
   - Phase Transition

Boltzmann Sequence-Structure Channel



$$P(f|s) := \frac{e^{-\beta\mathcal{E}(s,f)}}{Z(s,\beta)} \qquad Z(s,\beta) := \sum_{f\in\mathcal{F}} e^{-\beta\mathcal{E}(s,f)}$$

Sequences: $S = (S_1, \ldots S_N)$, i.i.d. with $P(S_i = H) = p = 1 - P(S_i = P)$.

β: a parameter that is meant to represent inverse temperature.

Folds: $\mathcal{F}_N$ denotes the set of self-avoiding walks of length N filling a square in $\mathbf{Z}^2$ of size N, starting at $(0,0)$ and ending at $(\sqrt{N}-1, \sqrt{N}-1)$.

Energy: $\mathcal{E}(s,f)$ denotes energy for a fold f computed as follows: for a given symmetric 2×2 scoring matrix $Q = \{Q_{ij}\}_{i,j\in\{1,2\}}$ define

$$\mathcal{E}(f,s) = 2(Q_{11}c_{HH} + Q_{22}c_{PP} + Q_{12}c_{HP}) =: X_i(f,s) \qquad (1)$$

where c_{xy} denotes the number of (non-adjacent) contacts in a fold.

Information Theoretic Quantities

Capacity :

$$C = \max_{P(S)} I(S;F) = \max_{P(S)} [H(F) - H(F|S)]$$

where **conditional entropy** is

$$H(F|S) = -\sum_{s\in\mathcal{S}_N} p(s) \sum_{f\in\mathcal{F}_N} p(f|s)\log p(f|s) = \mathbf{E}[\log Z(S,\beta)] + \beta \sum_{s,f} p(f,s)\mathcal{E}(f,s)$$

$$= \mathbf{E}[\log Z(S,\beta)] + \beta \mathbf{E}[\mathcal{E}_{\beta,S}(F)],$$

where the Boltzmann energy $\mathcal{E}_{\beta,S}(F)$ becomes:

$$\mathbf{E}[\mathcal{E}_{\beta,S}(F)] = \sum_{s,f} p(f,s)\mathcal{E}(f,s) = -\frac{d}{d\beta}\mathbf{E}[\log Z(S,\beta)].$$

Information Theoretic Quantities

Capacity :

$$C = \max_{P(S)} I(S;F) = \max_{P(S)} [H(F) - H(F|S)]$$

where **conditional entropy** is

$$H(F|S) = -\sum_{s\in\mathcal{S}_N} p(s) \sum_{f\in\mathcal{F}_N} p(f|s)\log p(f|s) = \mathbf{E}[\log Z(S,\beta)] + \beta \sum_{s,f} p(f,s)\mathcal{E}(f,s)$$

$$= \mathbf{E}[\log Z(S,\beta)] + \beta \mathbf{E}[\mathcal{E}_{\beta,S}(F)],$$

where the Boltzmann energy $\mathcal{E}_{\beta,S}(F)$ becomes:

$$\mathbf{E}[\mathcal{E}_{\beta,S}(F)] = \sum_{s,f} p(f,s)\mathcal{E}(f,s) = -\frac{d}{d\beta}\mathbf{E}[\log Z(S,\beta)].$$

This energy should be distinguished from $\mathbf{E}[\mathcal{E}(f,S)]$ for given f which is

$$\mathbf{E}[\mathcal{E}(f,S)] = \sum_i \mathbf{E}[X_i(f,S)] = N\alpha + O(\sqrt{N}) \geq \mathbf{E}[\mathcal{E}_{\beta,S}(F)]$$

since Boltzmann distribution gives higher probability to lower energy.

Information Theoretic Quantities

Capacity :

$$C = \max_{P(S)} I(S;F) = \max_{P(S)}[H(F) - H(F|S)]$$

where **conditional entropy** is

$$H(F|S) = -\sum_{s\in\mathcal{S}_N} p(s) \sum_{f\in\mathcal{F}_N} p(f|s)\log p(f|s) = \mathbf{E}[\log Z(S,\beta)] + \beta \sum_{s,f} p(f,s)\mathcal{E}(f,s)$$

$$= \mathbf{E}[\log Z(S,\beta)] + \beta \mathbf{E}[\mathcal{E}_{\beta,S}(F)],$$

where the Boltzmann energy $\mathcal{E}_{\beta,S}(F)$ becomes:

$$\mathbf{E}[\mathcal{E}_{\beta,S}(F)] = \sum_{s,f} p(f,s)\mathcal{E}(f,s) = -\frac{d}{d\beta}\mathbf{E}[\log Z(S,\beta)].$$

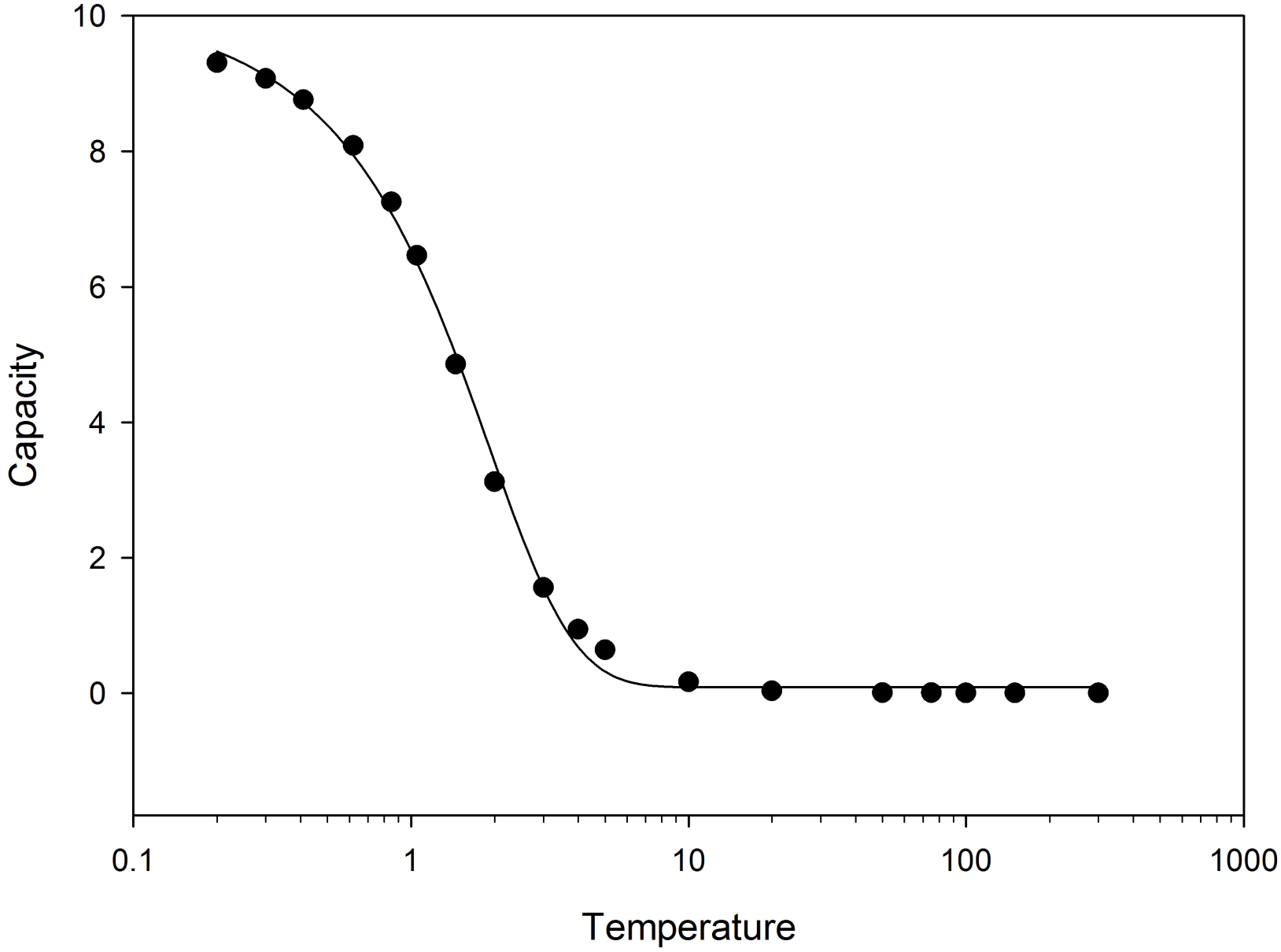
This energy should be distinguished from $\mathbf{E}[\mathcal{E}(f,S)]$ for given f which is

$$\mathbf{E}[\mathcal{E}(f,S)] = \sum_i \mathbf{E}[X_i(f,S)] = N\alpha + O(\sqrt{N}) \geq \mathbf{E}[\mathcal{E}_{\beta,S}(F)]$$

since Boltzmann distribution gives higher probability to lower energy.

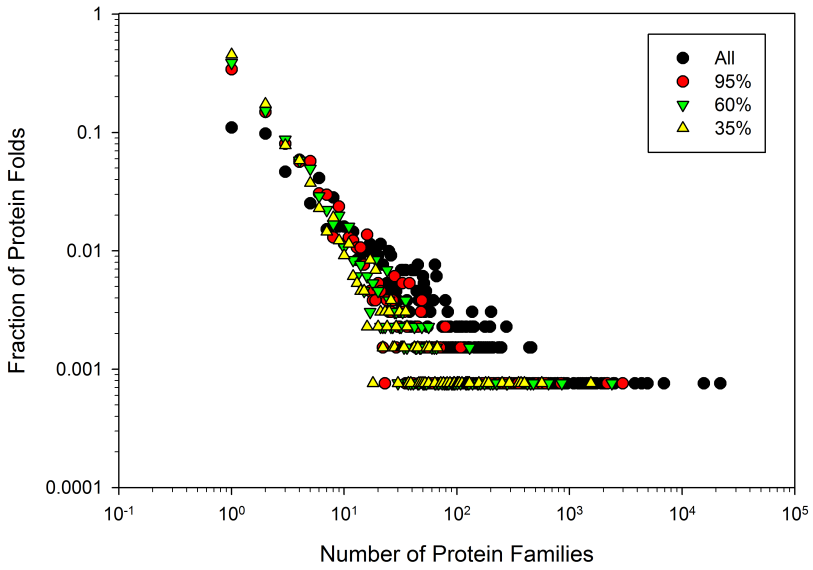
Example: For $Q = \begin{array}{c} \\ H \\ P \end{array}\begin{array}{c} \begin{array}{cc} H & P \end{array} \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array}$ we find $\mathbf{E}[\mathcal{E}(f,S)] = 2pqN + O(\sqrt{N}).$

Capacity Conjecture



Biological Motivation

Protein Folds in Nature

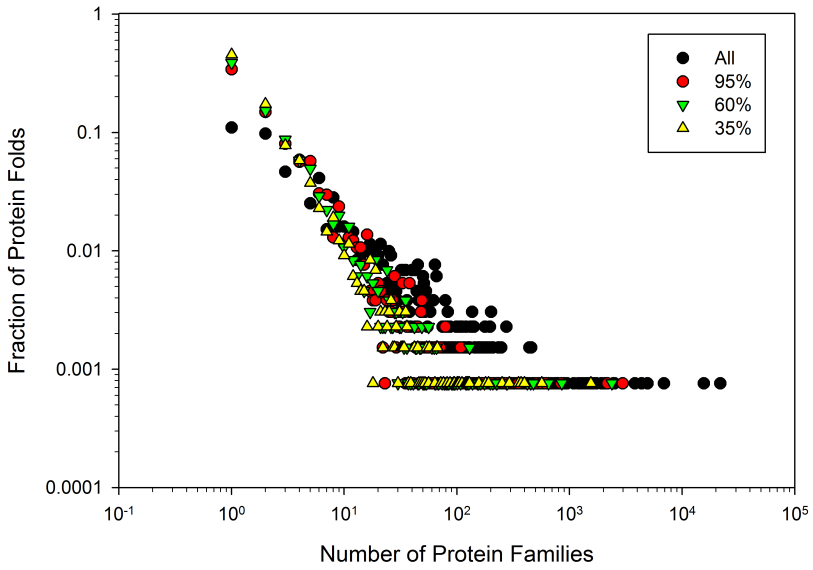


For each possible cardinality of protein families (x axis), count the number of protein folds (or sequences) observed in nature which are associated with that number of families. Plot on y axis the fraction of protein folds.
In nature, we observe lots of sequences with few associated folds and few sequences with lots of associated folds.

- The channel is a model of protein folding.
- Sequence distribution in nature exhibits a power law. In the channel model, such distributions (empirically) almost achieve capacity (nature prefers to avoid ambiguity!): capacity may have biological significance.

Biological Motivation

Protein Folds in Nature



For each possible cardinality of protein families (x axis), count the number of protein folds (or sequences) observed in nature which are associated with that number of families. Plot on y axis the fraction of protein folds.
In nature, we observe lots of sequences with few associated folds and few sequences with lots of associated folds.

- The channel is a model of protein folding.
- Sequence distribution in nature exhibits a power law. In the channel model, such distributions (empirically) almost achieve capacity (nature prefers to avoid ambiguity!): capacity may have biological significance.

Magner, Szpankowski, Kihara, "On the Origin of Protein Superfamilies and Superf olds", Scientific Reports, 2015.

Back to Theory . . .

Recall that

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

where we set

$$\alpha_*(\beta, N) = \alpha_*(\beta) = \frac{\mathbf{E}[\mathcal{E}_{\beta,S}(F)]}{N} = -\frac{d}{d\beta}\frac{\mathbf{E}[\log Z(S,\beta)]}{N}.$$

Furthermore, define free energy $\gamma(\beta, S)$ as

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

By submultiplicativity property of F_N we may conclude (needs a proof)

$$\limsup_{N\to\infty} \frac{\log |F_N|}{N} = \mu > 1.$$

Then

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

leading to

$$\boxed{H(F|S) \sim N[\gamma(\beta, S)\mu + \beta\alpha^*(\beta)].}$$

Main Results - Phase Transition

Theorem 8. *For any distribution over $\mathcal{S}_N$, $\beta > 0$, and scoring matrix Q satisfying a "niceness" condition, there exists $\sigma^2 > 0$ such that*

$$\mathbf{Var}[\mathcal{E}(f, S)] \sim N\sigma^2 > 0.$$

Then we have the following **phase transition**:

$$\limsup_{N\to\infty} \frac{H(F|S)}{N} \leq \begin{cases} \mu + \frac{1}{2}\sigma^2\beta^2 - \beta(\alpha - \alpha^*) & \beta > 0 \\ \beta\sqrt{2\sigma^2\mu} - \beta(\alpha - \alpha^*) & \beta \geq \beta_* = \frac{\sqrt{2\mu}}{\sigma}. \end{cases}$$

The conditional entropy phase transition is a consequence of the free energy phase transition:

$$\mu \cdot \gamma(\beta, S) \leq \begin{cases} \mu - \beta\alpha + \frac{1}{2}\sigma^2\beta^2 & \beta < \frac{\sqrt{2\mu}}{\sigma} \\ \beta(\sqrt{2\sigma^2 \log \mu} - \alpha) & \beta \geq \frac{\sqrt{2\mu}}{\sigma} \end{cases}$$

Remark. There is an information-theoretic upper bound on $H(F|S)$:

$$H(F|S) \leq H(F) \leq \log |\mathcal{F}_N| = N\mu$$

which may beat the upper bound above since we know that $\alpha - \alpha^* > 0$!

Matching Lower Bound?

Finding a lower matching bound is very challenging. It depends on the covariance between energies of two folds f and g.

Let K denote the number of shared contacts between two folds f, g chosen uniformly at random with replacement.

Theorem 9 (Free energy lower bound for high temperature)**.** *There exists a scoring matrix for which we have for large* N

$$\frac{H(F|S)}{N} \geq \mu - \beta(\alpha - \alpha_*(\beta)) + \frac{1}{2}\beta^2\sigma^2 - o(1),$$

provided $\beta = o(1/\sqrt{N \log N})$.

Remark. For Q satisfying Theorem 9 and $\beta = o(1/\sqrt{N \log N})$ we observe that

$$\alpha - \alpha_* \sim \beta\sigma^2$$

leading to

$$H(F|S) \sim N\left(\mu_N - \frac{1}{2}\sigma^2\beta^2\right)$$

for small $\beta < \beta_*$.

Back to Capacity Conjecture and CounterExample

Corollary 1. *With p and the scoring matrix Q are as above, and $\beta_N = o(N^{-2/3}\log^{-1/3} N)$, then*

$$I(F;S) = o(1).$$

Back to Capacity Conjecture and CounterExample

Corollary 1. *With p and the scoring matrix Q are as above, and $\beta_N = o(N^{-2/3}\log^{-1/3} N)$, then*

$$I(F;S) = o(1).$$

There are scoring matrices for which there is no phase transition, as shown in the next theorem.

Theorem 10. *Let Q be the scoring matrix which maps:*

$$HH \mapsto -1/2, \quad HP/PH \mapsto -1/4, \quad PP \mapsto 0.$$

Then, for arbitrary sequence distributions, the free energy is given by

$$\gamma(\beta) = 1 + \beta \limsup_{N\to\infty} \frac{\mathbf{E}[D_S(H)]}{\log|\mathcal{F}_N|},$$

where $D_S(H)$ is the number of i for which $S_i = H$. In the case of Bernoulli $S \sim \mathcal{B}_N(p)$, this becomes

$$\gamma(\beta) = 1 + \frac{\beta p}{\mu} = 1 - \frac{\beta\alpha}{\mu}.$$

That is, there is no phase transition!

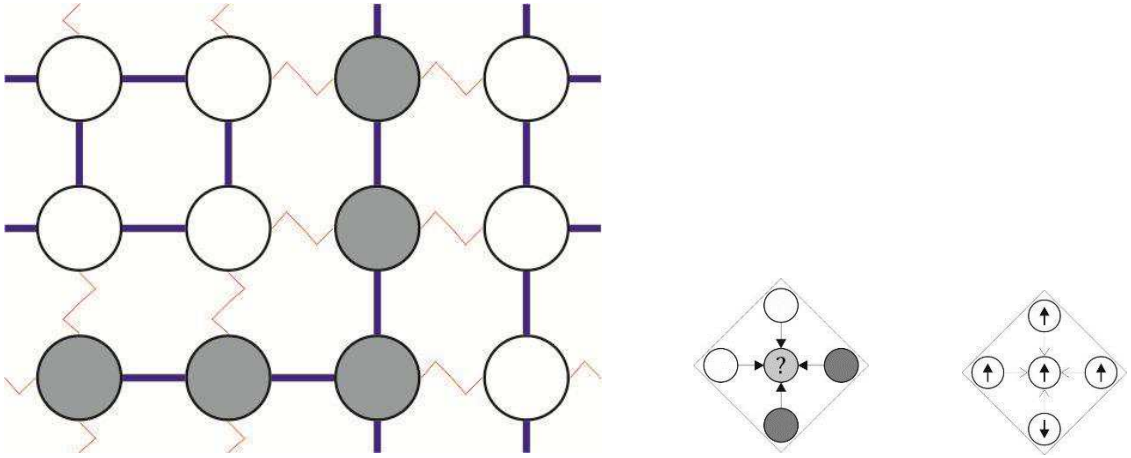
That's It

THANK YOU

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 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.

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 Channel

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

One-Dimensional Markov: Sequences $x^n = x_1 \dots x_n$ over $\mathcal{A} = \{1, 2, \dots, m\}$ 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

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

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 $x^n = x_1 \dots x_n$ over $\mathcal{A} = \{1, 2, \dots, m\}$ 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

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

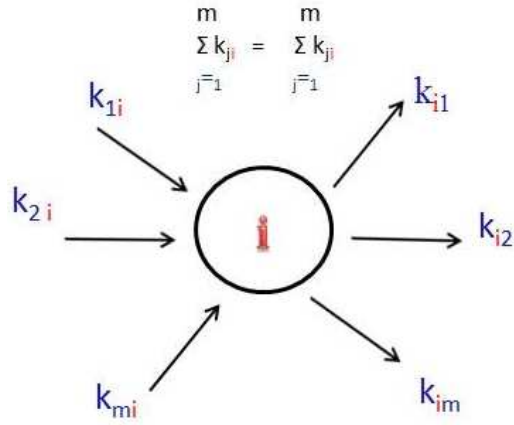
where k_{ij} is the number of pair symbols (ij) in x_1^n, that is, i followed by j.

For circular strings (i.e., after the nth symbol we re-visit the first symbol of x_1^n), the matrix $\mathbf{k} = [k_{ij}]$ satisfies the following constraints denoted as $\boxed{\mathcal{F}_n(m)}$:

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

For example: m=3

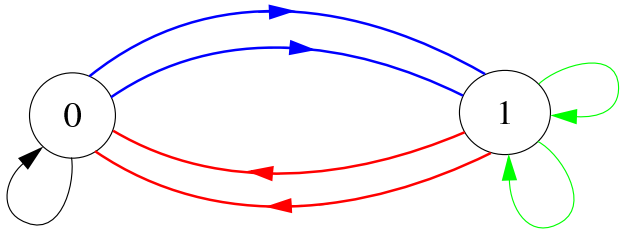
$k_{11} + k_{12} + k_{13} + k_{21} + k_{22} + k_{23} + k_{31} + k_{32} + k_{33} = n$
$k_{12} + k_{13} = k_{21} + k_{31}$
$k_{12} + k_{32} = k_{21} + k_{23}$
$k_{13} + k_{23} = k_{31} + k_{32}$



Markov Types and Eulerian Cycles

Example: Let $\mathcal{A} = \{0, 1\}$ and

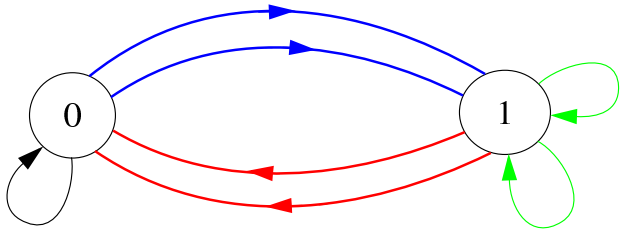
$$\mathbf{k} = \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$$



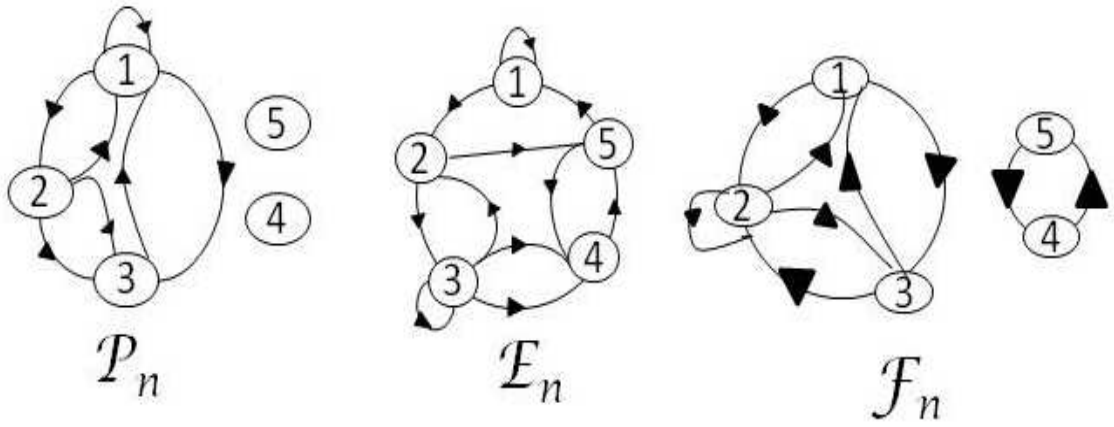
Markov Types and Eulerian Cycles

Example: Let $\mathcal{A} = \{0, 1\}$ and

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



$\mathcal{P}_n(m)$ – 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 11. (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

$$d(m) = \frac{1}{(2\pi)^{m-1}} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{(m-1)-fold} \prod_{j=1}^{m-1} \frac{1}{1+\varphi_j^2} \cdot \prod_{k\neq\ell} \frac{1}{1+(\varphi_k-\varphi_\ell)^2}\, d\varphi_1 d\varphi_2 \cdots d\varphi_{m-1}.$$

(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 $1.767043356\ 10^{-11}$.

Outline Update

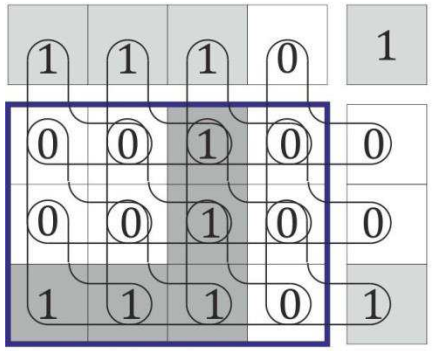
1. Structural Compression
2. Tree Compression
   - One Dimensional Markov Types
   - One-Dimensional Universal Types
   - Markov Fields and Tilings
3. Sequence-Structure Channel

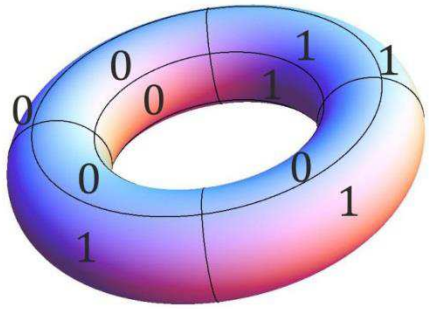
Outline Update

1. Structural Compression
2. Structure of Markov Fields
   - One Dimensional Markov Types
   - Markov Fields and Tilings
3. Sequence-Structure 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 $\mathcal{O}_{\mathbf{n}}$.
The shape of interaction is $S \subset \mathbb{Z}^d$.
A tile t is $t : S \to \mathcal{A}$ and $T = \{t : S \to \mathcal{A}\}$.





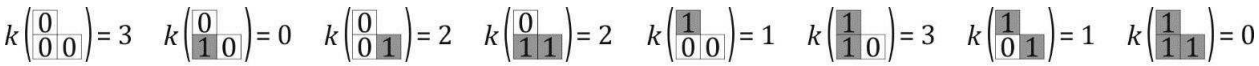

shape


tile

Markov Field Type $\mathcal{X}^{\mathbf{n}} = \{x^{\mathbf{n}} : \mathcal{O}_{\mathbf{n}} \to \mathcal{A}\}$:
Define the frequency vector of dimension $D = |T| = m^{|S|}$:

$$k(t) \equiv k_S(t) = |\{\mathbf{s} \in \mathcal{O}_{\mathbf{n}} : x|_{S+\mathbf{s}} = t\}|, \quad t \in T.$$

Example:



3 + 0 + 2 + 2 + 1 + 3 + 1 + 0 = 12 size of torus

A set of Markov field types or tile types is:

$$\mathcal{P}_{\mathbf{n}}(m, S) = \{\mathbf{k} : \exists_{x \in \mathcal{X}_{\mathbf{n}}} x^{\mathbf{n}} \text{is of type } \mathbf{k}\}.$$

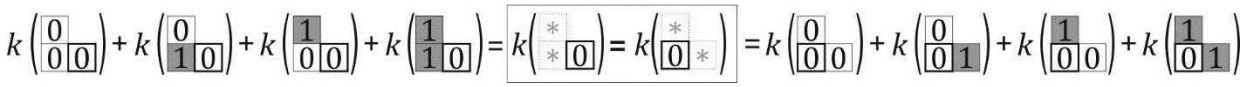
Conservation Laws

Conservation Laws:

$$\forall_{\emptyset \neq S' \subset S,\ \mathbf{s} \in \mathbb{Z}^d : (S'+\mathbf{s}) \subset S}\ \forall_{t' : S' \to \mathcal{A}} \quad k_{S'}(t') = k_{S'+\mathbf{s}}(t')$$

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

Example:



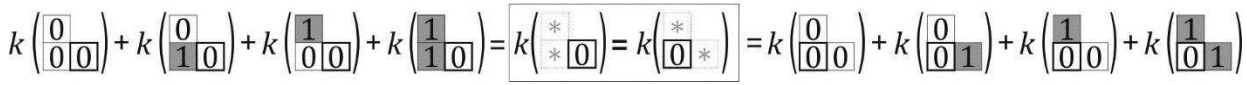
Conservation Laws

Conservation Laws:

$\forall_{\emptyset \neq S' \subset S,\ \mathbf{s} \in \mathbb{Z}^d : (S'+\mathbf{s}) \subset S}\ \forall_{t' : S' \to \mathcal{A}} \quad k_{S'}(t') = k_{S'+\mathbf{s}}(t')$

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

Example:



The conservation laws can be viewed as *linear equations* with a $1 \times D$ row denoted as $C(\{(S', \mathbf{s}, t')\})$.
The matrix C^* is hugely over determined! Our goal is to find C such that the conservation laws can be written as

$$C\mathbf{k} = \mathbf{0}.$$

Example 1. $d = 1$-*dimensional Markov over* $\mathcal{A} = \{1, 2\}$.
Tiles are $((11), (21), (12), (22))$ and the conservation laws are

$$\begin{aligned} k(11) + k(12) &= k(1*) = k(*1) = k(11) + k(21), \\ k(21) + k(22) &= k(2*) = k(*2) = k(12) + k(22). \end{aligned}$$

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

$$(0, -1, 1, 0) \cdot \mathbf{k} = \mathbf{0}.$$

Dimension of the Frequency Vectors

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

$$\boxed{\mu = D - \mathrm{rk}(C) - 1.}$$

where $\mathrm{rk}(C)$ is the rank of matrix C. I

Theorem 12. *The matrix C has rank*

$$\mathrm{rk}(C) = \sum_{S' \in \mathcal{S}^0} (|\{\mathbf{s} : (S' + \mathbf{s}) \subset S\}| - 1)(m-1)^{|S'|}$$

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

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

Dimension of the Frequency Vectors

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

$$\boxed{\mu = D - \mathrm{rk}(C) - 1.}$$

where $\mathrm{rk}(C)$ is the rank of matrix C. I

Theorem 12. *The matrix C has rank*

$$\mathrm{rk}(C) = \sum_{S' \in \mathcal{S}^0} (|\{\mathbf{s} : (S' + \mathbf{s}) \subset S\}| - 1)(m-1)^{|S'|}$$

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

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

Example: For $d = 2$ and a 2×2 square shape we have $\mu = m^4 - 2m^2 + m$, while for a 3×2 rectangular shape we find $\mu = m^6 - m^4 - m^3 + m^2$.

Geometry

We view the count vector $\mathbf{k} = \{k(t)\}_{t \in T}$ in the $D = m^{|S|}$ space.

$$\mathscr{C} = \{\mathbf{k} \in \mathbb{N}^D \ : \ C_m(S) \cdot \mathbf{k} = \mathbf{0}\}$$
$$\mathcal{F} = \{\mathbf{k} \in \mathscr{C} : \textstyle\sum_i k_i = N\}$$
$$\hat{\mathcal{F}} = \{\hat{\mathbf{k}} \in \{(\mathbb{R}_+^D \ : \ C \cdot \hat{\mathbf{k}} = \mathbf{0}, \ \textstyle\sum_i \hat{k}_i = 1\}$$
$$\mathcal{F}_N = \{N\hat{\mathbf{k}} : \hat{\mathbf{k}} \in \hat{\mathcal{F}}, \ N\hat{\mathbf{k}} \in \mathbb{Z}^D\}$$

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

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

Geometry

We view the count vector $\mathbf{k} = \{k(t)\}_{t \in T}$ in the $D = m^{|S|}$ space.

$$\mathscr{C} = \{\mathbf{k} \in \mathbb{N}^D \ : \ C_m(S) \cdot \mathbf{k} = \mathbf{0}\}$$
$$\mathcal{F} = \{\mathbf{k} \in \mathscr{C} : \textstyle\sum_i k_i = N\}$$
$$\hat{\mathcal{F}} = \{\hat{\mathbf{k}} \in \{(\mathbb{R}_+^D \ : \ C \cdot \hat{\mathbf{k}} = \mathbf{0}, \ \textstyle\sum_i \hat{k}_i = 1\}$$
$$\mathcal{F}_N = \{N\hat{\mathbf{k}} : \hat{\mathbf{k}} \in \hat{\mathcal{F}}, \ N\hat{\mathbf{k}} \in \mathbb{Z}^D\}$$

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

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

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

Theorem 13 (Ehrhart, 1967). *If $\hat{\mathcal{F}}$ 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*

$$|\mathcal{F}_N| = a_{\mu,j} N^\mu + a_{\mu-1,j} N^{\mu-1} + \ldots a_{0,j} \quad N \equiv j \pmod{p}.$$

Main Result for Markov Types

Theorem 14. *Consider the torus* $\mathcal{O}_{\mathbf{n}}$. *There exists* $0 < c^{min} \leq c^{max}$ *such that*

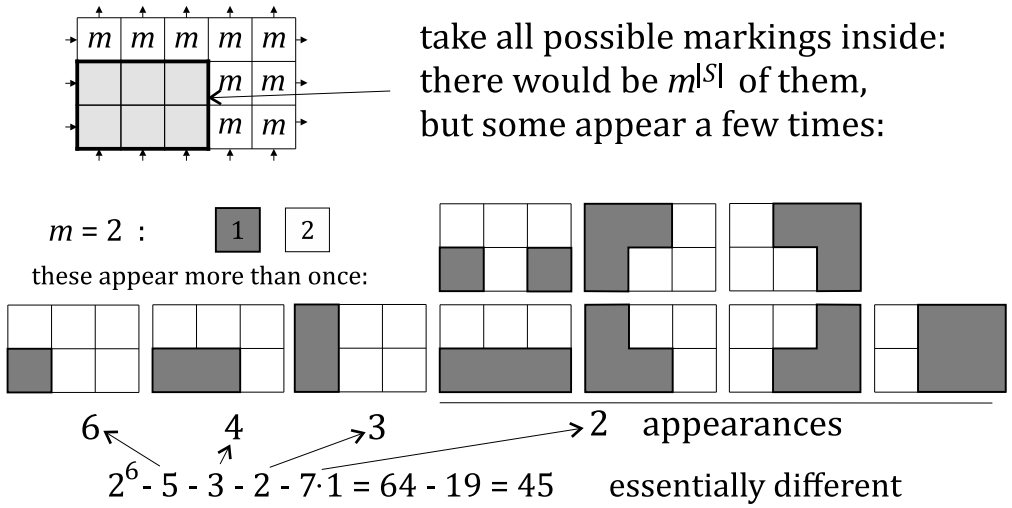
$$c^{min} N^{\mu} \leq |\mathcal{P}_{\mathbf{n}}(m, S)| \leq c^{max} N^{\mu}$$

Main Result for Markov Types

Theorem 14. *Consider the torus* $\mathcal{O}_{\mathbf{n}}$. *There exists* $0 < c^{min} \leq c^{max}$ *such that*

$$c^{min} N^{\mu} \leq |\mathcal{P}_{\mathbf{n}}(m, S)| \leq c^{max} N^{\mu}$$

Lemma 4. *There exist* $\mu + 1$ *linearly independent periodic tilings.*

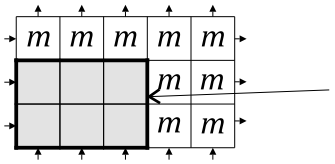


Main Result for Markov Types

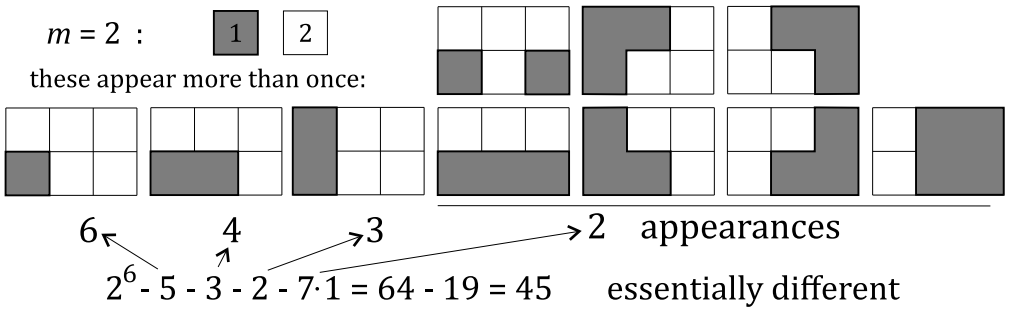
Theorem 14. *Consider the torus* $\mathcal{O}_{\mathbf{n}}$. *There exists* $0 < c^{min} \leq c^{max}$ *such that*

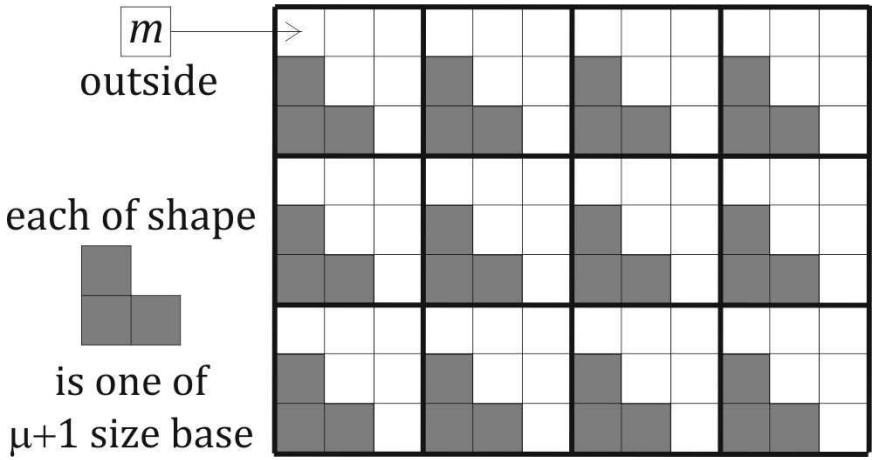
$$c^{min} N^{\mu} \leq |\mathcal{P}_{\mathbf{n}}(m, S)| \leq c^{max} N^{\mu}$$

Lemma 4. *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:





$$\frac{a_1 \hat{\mathbf{k}}^1 + .. + a_{\mu+1} \hat{\mathbf{k}}^{\mu+1}}{a_1 + \cdots + a_{\mu+1}} : \forall_i \, a_i \in \mathbb{N}, \; \sum_i a_i = N$$

$$\binom{N+\mu-1}{\mu} = O(N^{\mu})$$

That's It

THANK YOU