Entropy of Some Advanced Data Structures

Wojciech Szpankowski
Purdue University
W. Lafayette, IN 47907

June 12, 2017

AofA, Princeton, 2017

---

*Joint work with Y. Choi, Z. Golebiewski, S. Janson, T. Luczak, A. Magner, and K. Turowski.

Outline

1. Multimodal and Multi-context Data Structures
2. Entropy of Binary Trees
   - Motivation
   - Plane vs Non-Plane Trees
   - Entropy Computation
3. Entropy of General d-ary Trees
   - m-ary Search Trees
   - d-ary Trees
   - General Trees
4. Entropy of Graphs
   - Structural Entropy – Unlabeled Graphs
   - Erdős–Renyí Graphs
   - Preferential Attachment Graphs

Multimodal Data Structures

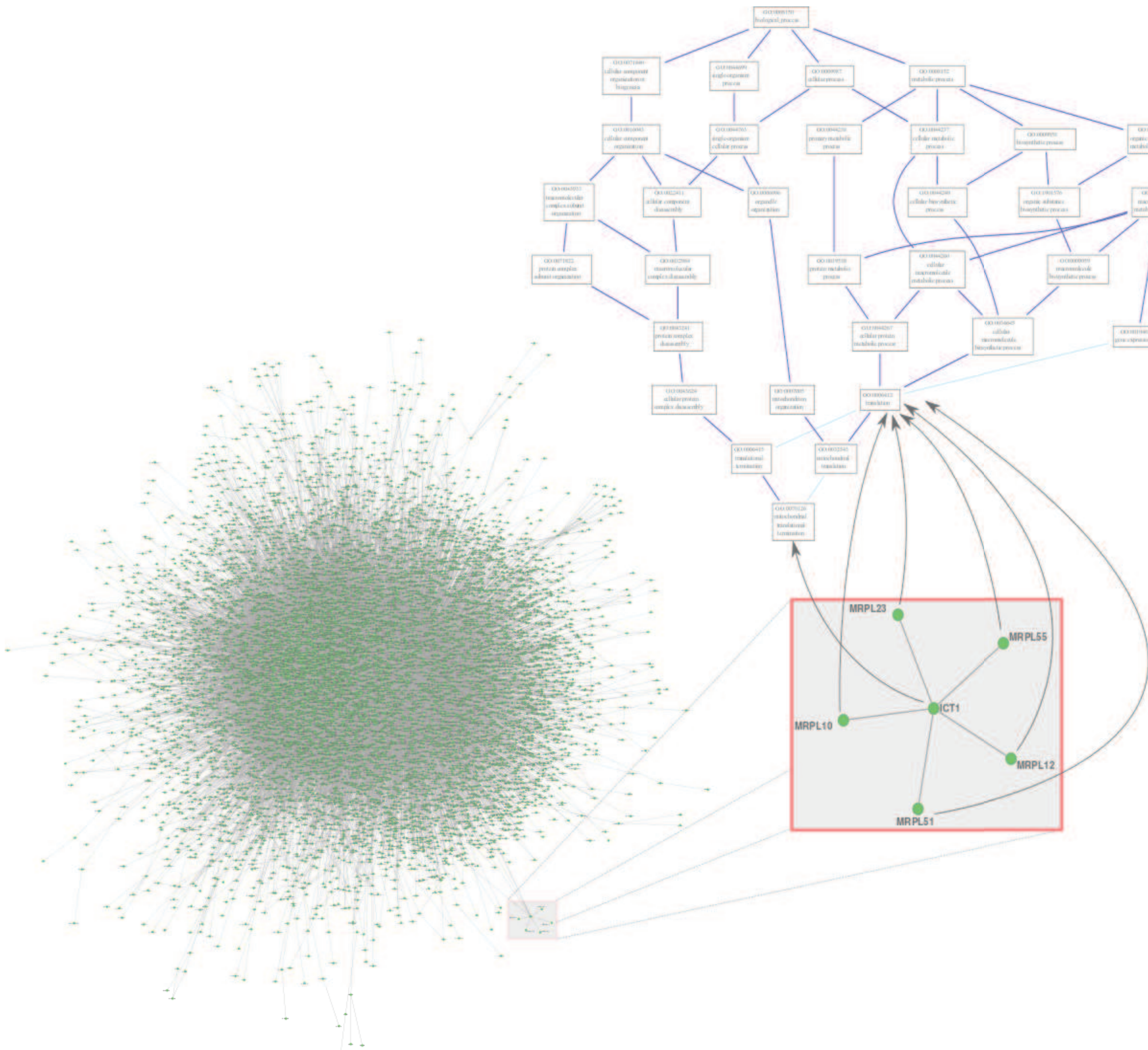


Figure 1: Protein-Protein Interaction Network with BioGRID database

Outline Update

1. Multimodal and Multi-context Data Structures
2. Entropy of Binary Trees
   - Motivation
   - Plane vs Non-Plane Trees
   - Entropy Computation
3. Entropy of General d-ary Trees
4. Entropy of Graphs

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 $t = 1, ..., n$: Choose a leaf uniformly at random and attach 2 children.

Source models for trees

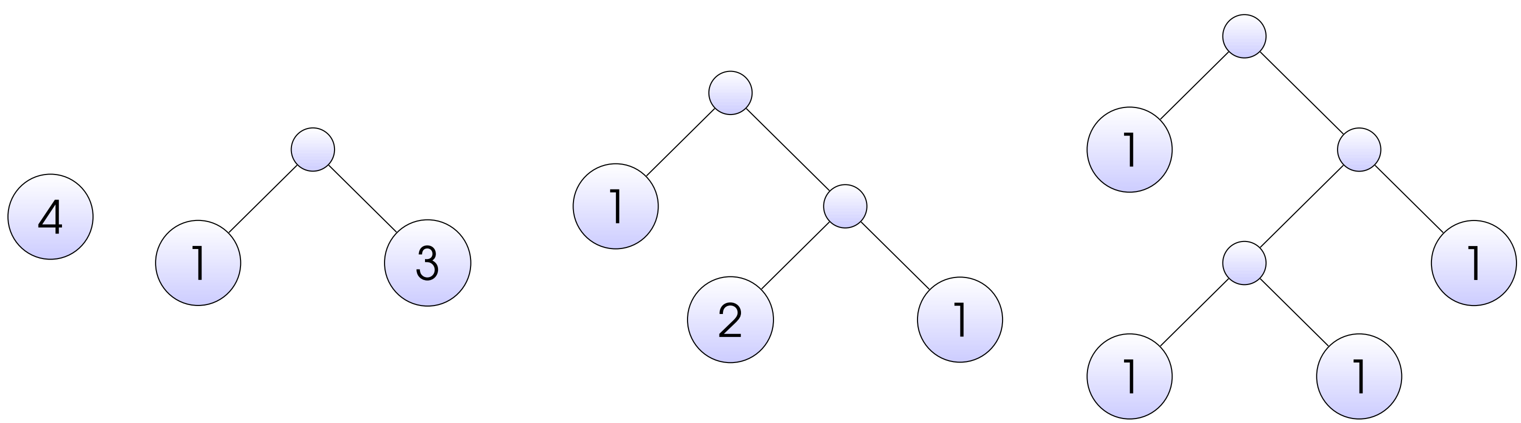
Probabilistic models for rooted binary plane trees:

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 (binary search tree):

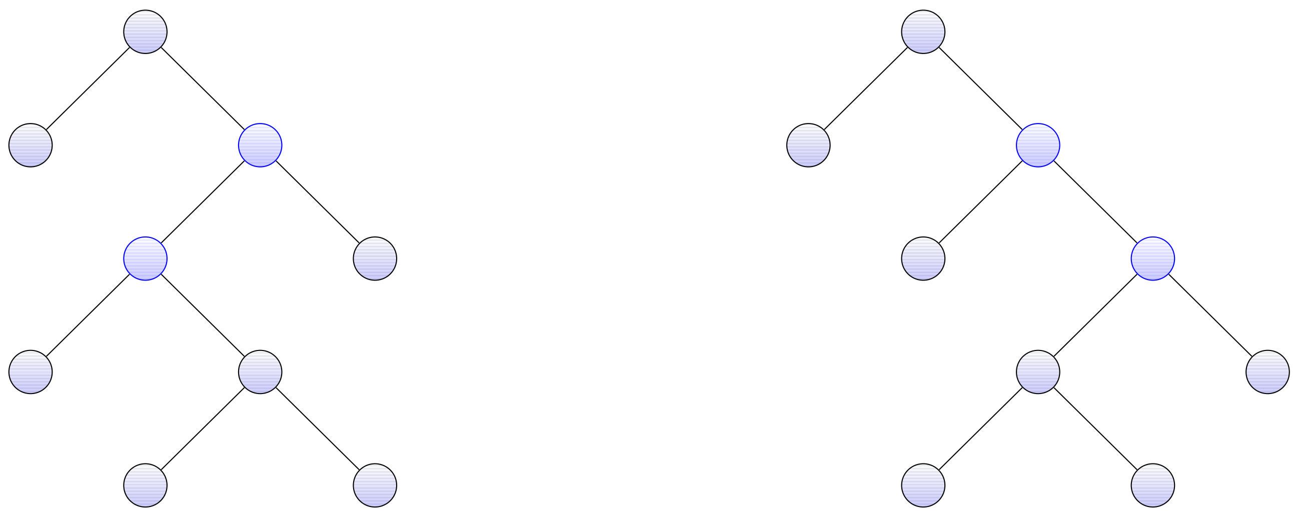
- Initially, add a node with label n.
- While there is a leaf with label $\ell > 1$, choose a number ℓ' uniformly at random from $[\ell - 1]$ and add a left and right child with labels with ℓ' and $\ell - \ell'$, respectively.



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.

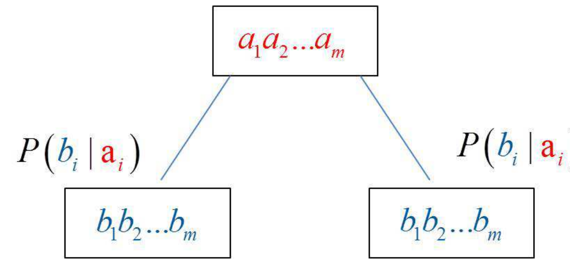
Example of two equivalent trees:



Source models for vertex names

Parameters for vertex names:

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



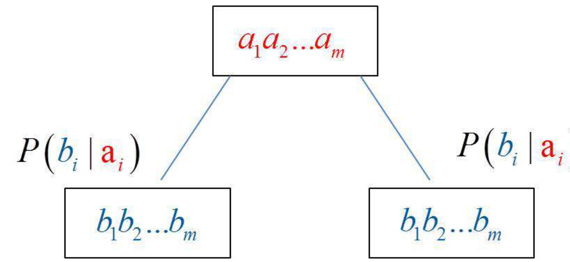
Generating vertex names given a tree structure:

- Generate a name for the root by taking m letters from a memoryless source with distribution π on $\mathcal{A}$.
- Given a name $a_1a_2, \ldots, a_m$ 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.

Source models for vertex names

Parameters for vertex names:

$\mathcal{A}$: The (finite) alphabet.
$m \geq 0$: the length of a name.
P: Markov transition matrix
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.

Tree Entropy:

$$H(T_n) = -\mathbf{E}[\log P(T_n)] = -\sum_{t_n \in T_n} P(T_n = t_n) \log P(T_n = t_n).$$

Entropy for Plane-Oriented Trees with Names

Theorem 1 (Magner, W.S., Turowski, 2016). *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.

See also Kieffer, Yang, W.S., ISIT 2009.

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 1. *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 2 (Magner, Turowski, W.S., 2016). *Entropy rate for non-plane trees is*

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

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 $b_1 = 1$ (see Hwang, Martinez, et al., 2012).

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

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

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

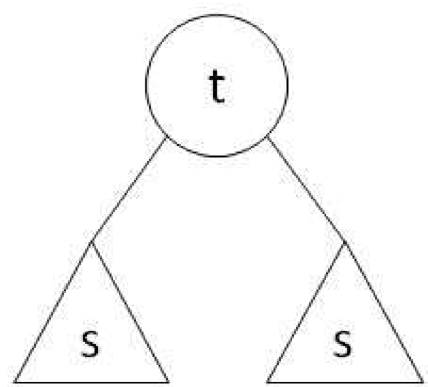
$$H(T_n|S_n) = -\sum_{t \in \mathcal{T}_n, s \in \mathcal{S}_n} \Pr(T_n = t, S_n = s) \log \Pr(T_n = t|S_n = s) = \mathbf{E}X_n + \mathbf{E}Y_n$$

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

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

where

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



Outline Update

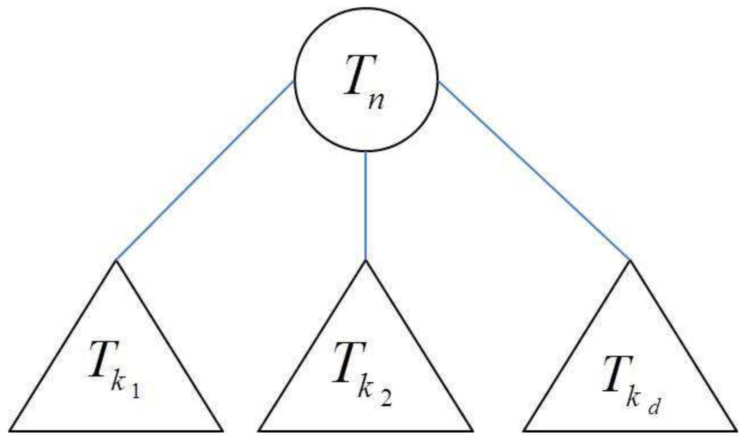
1. Multimodal and Multi-context Data Structures
2. Entropy of Binary Trees
3. Entropy of General d-ary Trees
   - m-ary Search Trees
   - d-ary Trees
   - General Trees
4. Entropy of Graphs

Generalized Trees

Let T_n represent a random tree t_n on n internal nodes.
No correlated names.

General Probabilistic Model:
Tree t_n is split into d subtrees of size $k_1, \ldots, k_d$ where

$$k_1 + \cdots + k_d = n - 1.$$



Then we assume that

$$P(T_n = t_n) = P(k_1, \ldots, k_d) \prod_{i=1}^{d} P\left(T_{k_i} = t_{k_i}\right)$$

where the **split probability** $P(k_1, \ldots, k_d)$ is the probability of the split at the root of sizes $k_1, \ldots k_d$, respectively.

This split probability $P(k_1, \ldots, k_d)$ is different for variety d-ary trees.

m-ary Search Trees

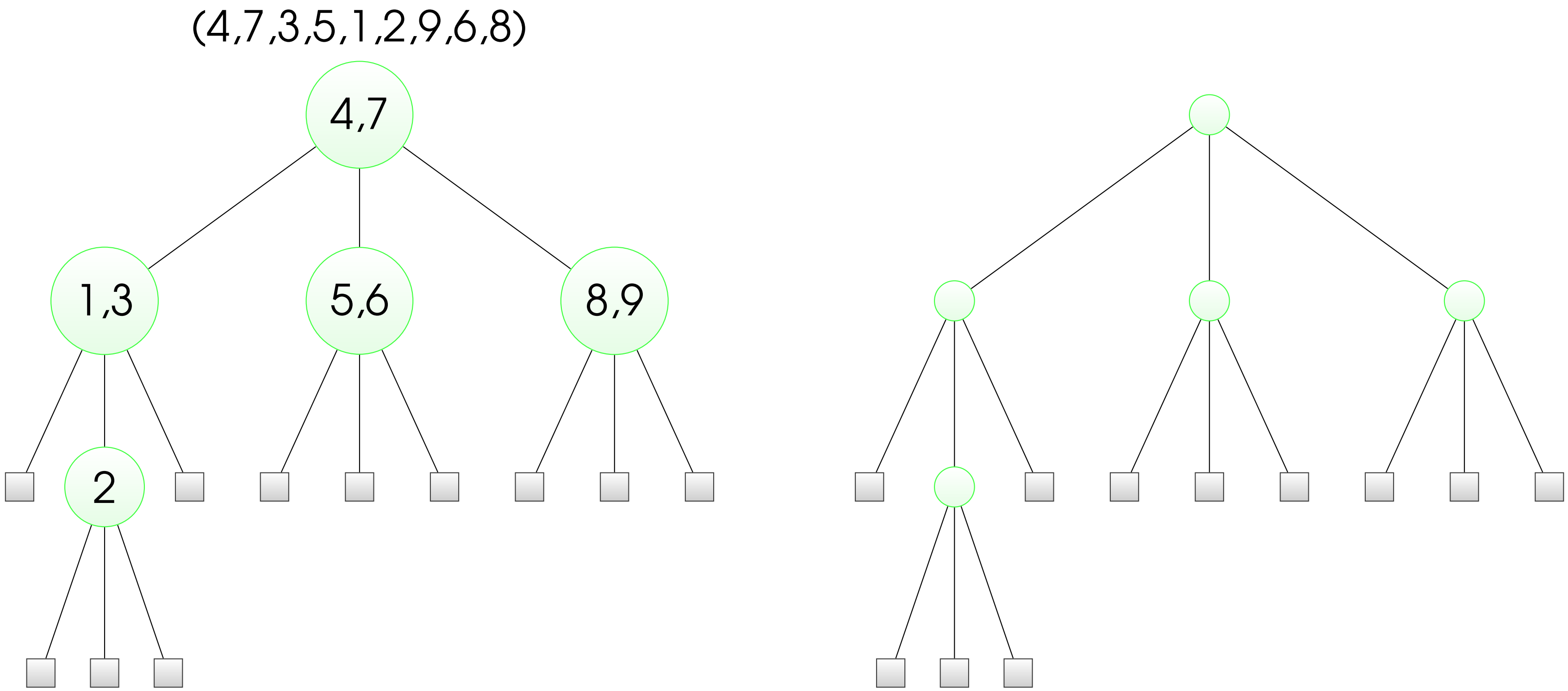


Figure 2: 3-ary search tree built over $(4, 7, 3, 5, 1, 2, 9, 6, 8)$.

m-ary Search Trees

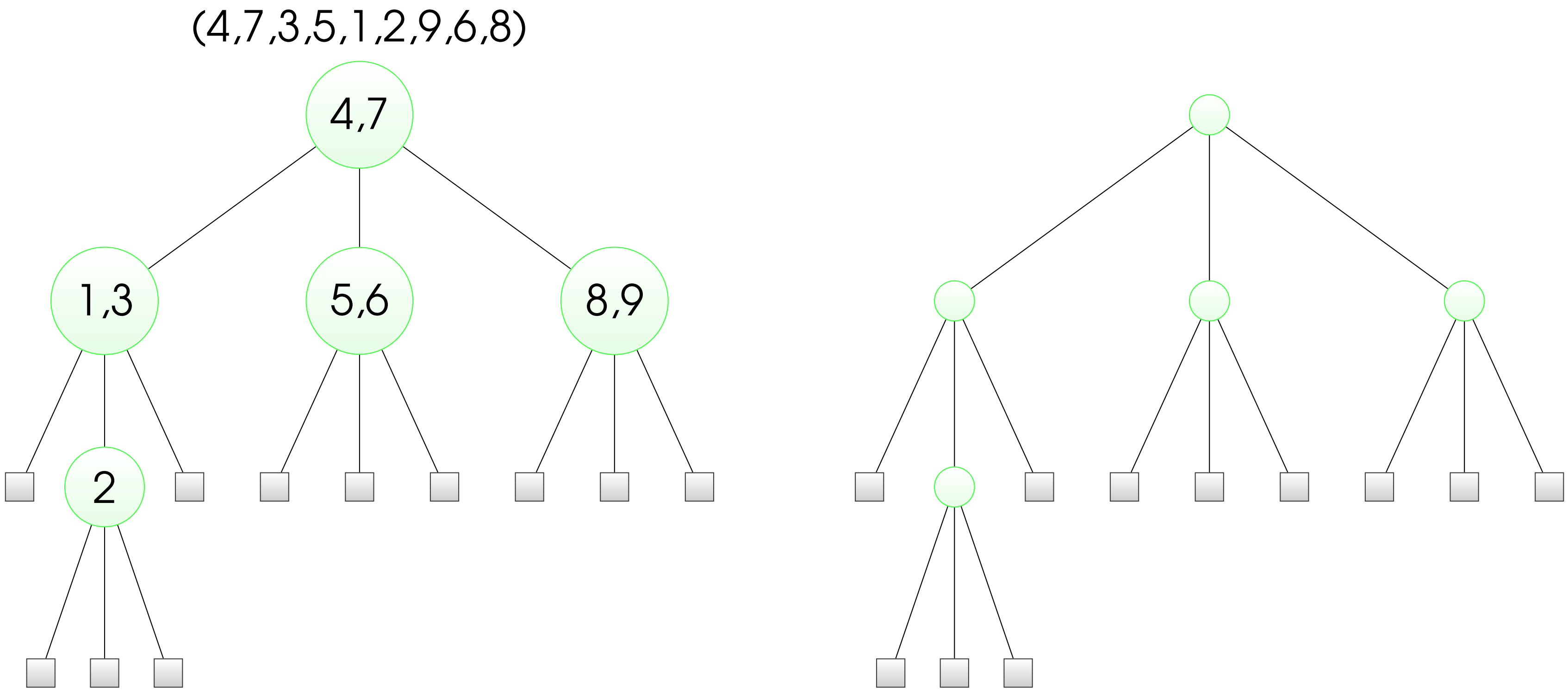


Figure 2: 3-ary search tree built over $(4, 7, 3, 5, 1, 2, 9, 6, 8)$.

Theorem 3 (Fill, Hwang, et al., 2005)**.** *For m-ary search tree, the entropy $H_n^{(m)}$ is*

$$H_n^{(m)} = n \cdot \frac{2}{2\mathcal{H}_m - 2} \sum_{k \geq 0} \frac{\log \binom{k}{m-1}}{(k+1)(k+2)} + o(n)$$

where $\mathcal{H}_m = \sum_{i=1}^{m} \frac{1}{i}$ is the harmonic number.

Generating d-ary Trees: 3-ary Example

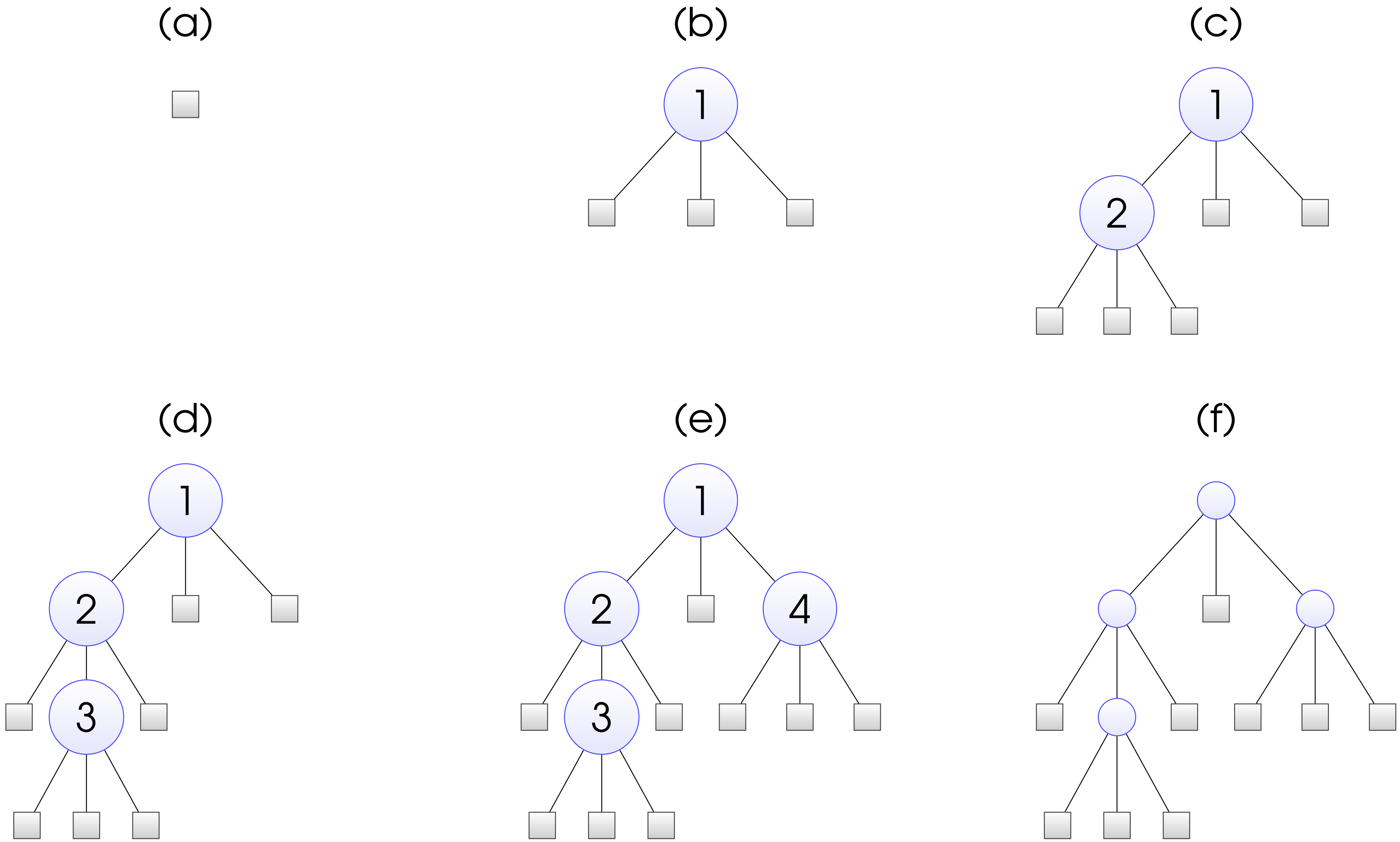


Figure 3: Labeled and unlabeled 3-ary trees of size 4.

General Plane Trees

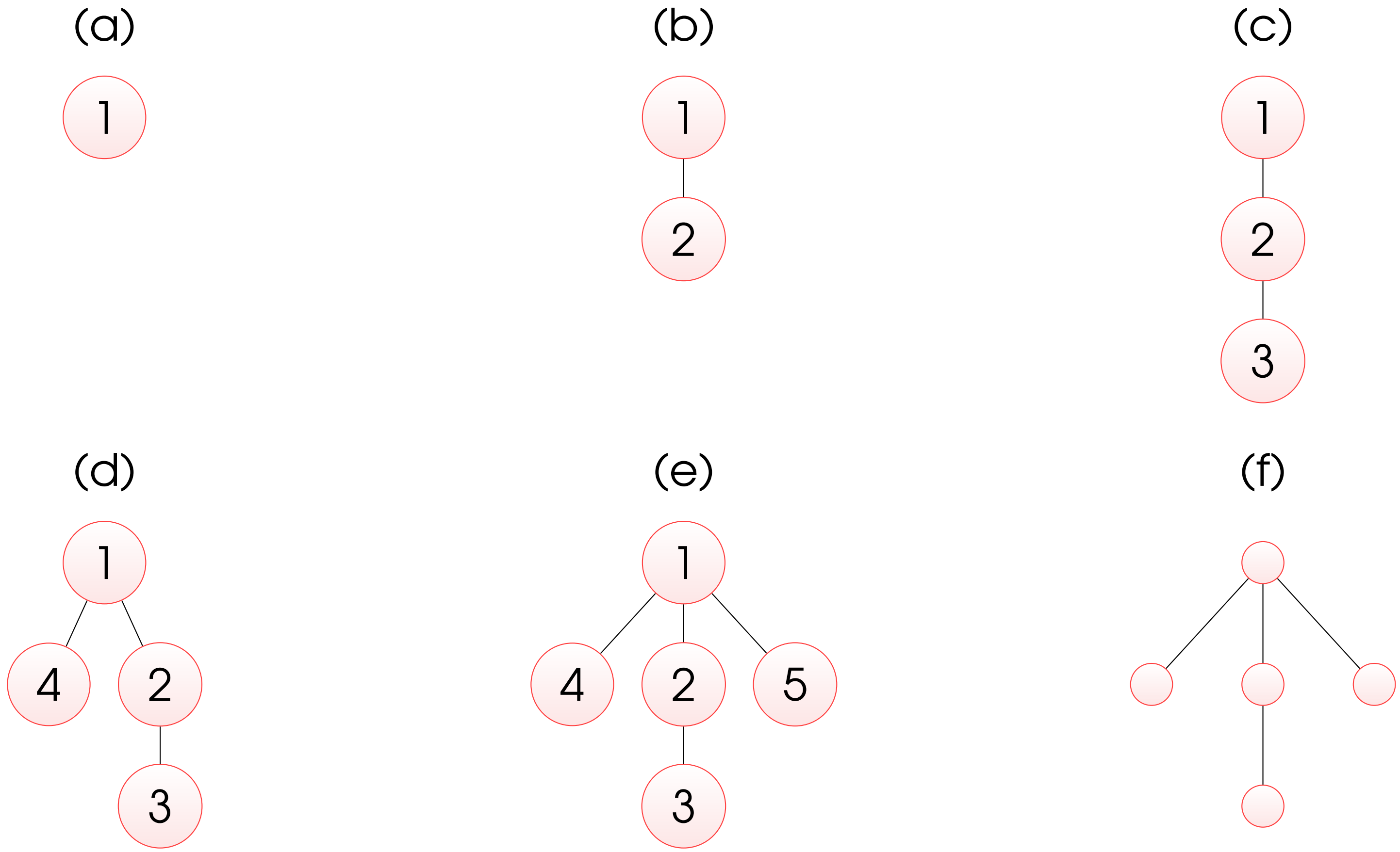


Figure 4: Labeled and unlabeled general trees of size 5.

Entropy of d-ary Trees

Let W_n denote d-ary trees with n internal nodes.

Entropy Recurrence:
We write $\mathbf{k} = (k_1, \ldots, k_d)$. Then

$$H(W_n) = H(P(\mathbf{k})) + d \sum_{k=0}^{n-1} p_{n,k} H(W_k)$$

where $P(\mathbf{k})$ is the split probability that can be expressed as

$$P(\mathbf{k}) = \binom{n-1}{k_1, \ldots, k_d} \frac{g_{k_1} \cdots g_{k_d}}{g_n}, \qquad g_n = (-1)^n (d-1)^n \frac{\Gamma(2 - \frac{d}{d-1})}{\Gamma(2 - \frac{d}{d-1} - n)}.$$

Let $\alpha = d/(d-1)$. Then $p_{n,k}$ is the prob. of one subtree of size k:

$$p_{n,k} = \sum_{k_2 + \cdots k_d = n-k} P(\mathbf{k}) = \frac{(\alpha-1)}{n} \frac{n!\Gamma(k+\alpha-1)}{k!\Gamma(n+\alpha-1)}.$$

Entropy of d-ary Trees

Let W_n denote d-ary trees with n internal nodes.

Entropy Recurrence:
We write $\mathbf{k} = (k_1, \ldots, k_d)$. Then

$$H(W_n) = H(P(\mathbf{k})) + d \sum_{k=0}^{n-1} p_{n,k} H(W_k)$$

where $P(\mathbf{k})$ is the split probability that can be expressed as

$$P(\mathbf{k}) = \binom{n-1}{k_1, \ldots, k_d} \frac{g_{k_1} \cdots g_{k_d}}{g_n}, \qquad g_n = (-1)^n (d-1)^n \frac{\Gamma(2 - \frac{d}{d-1})}{\Gamma(2 - \frac{d}{d-1} - n)}.$$

Let $\alpha = d/(d-1)$. Then $p_{n,k}$ is the prob. of one subtree of size k:

$$p_{n,k} = \sum_{k_2 + \cdots k_d = n-k} P(\mathbf{k}) = \frac{(\alpha - 1)}{n} \frac{n!\Gamma(k + \alpha - 1)}{k!\Gamma(n + \alpha - 1)}.$$

Theorem 4 (Golebiewski, Magner, W.S., 2017). *The entropy of a d-ary tree is*

$$H(W_n) = H(P(\mathbf{k})) + \alpha(n + \alpha - 1) \sum_{k=0}^{n-1} \frac{H(P(\mathbf{k}))}{(k + \alpha - 1)(k + \alpha)}$$

where $H(P(\mathbf{k}))$ is the entropy of $P(\mathbf{k})$.

Sketch of the Proof

Apply the following lemma with $a_n = H(P(\mathbf{k}))$.

Lemma 2. *For constant* α, x_0 *and* x_1*, the recurrence*

$$x_n = a_n + \frac{\alpha}{n}\frac{n!}{\Gamma(n+\alpha-1)}\sum_{k=0}^{n-1}\frac{\Gamma(k+\alpha-1)}{k!}x_k, \qquad n \geq 2$$

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

$$x_n = a_n + \alpha(n+\alpha-1)\sum_{k=0}^{n-1}\frac{a_k}{(k+\alpha-1)(k+\alpha)} + \frac{n+\alpha-1}{\alpha+1}\left(x_1 + \frac{x_0}{\alpha-1}\right).$$

Remark. For example for $d = 3$ we have

$$H(P(\mathbf{k})) = \log\left(\frac{n}{2^n}\binom{2n}{n}\right) - \frac{3}{2n}\sum_{k=0}^{n-1}\frac{\binom{2k}{k}2^{2n}}{\binom{2n}{n}2^{2k}}\log\left(\frac{\binom{2k}{k}}{2^k}\right)$$

and $H(W_n) \approx n \cdot 2.470$.

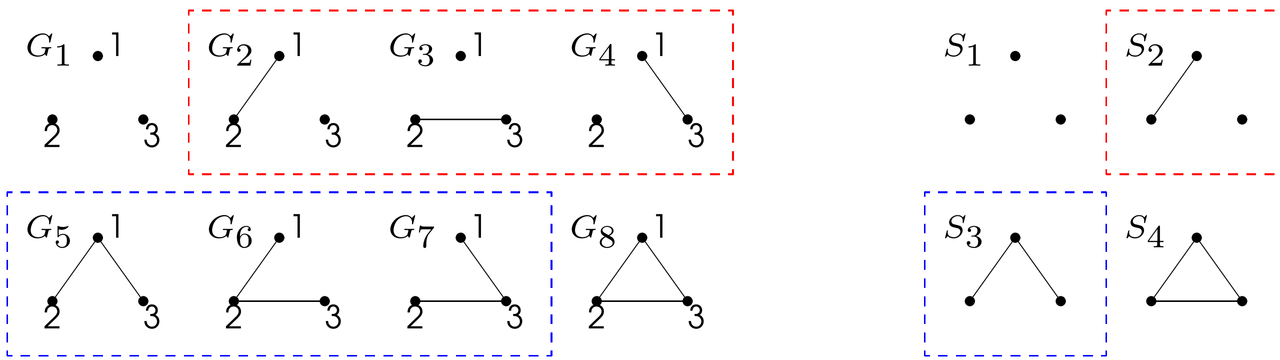
Outline Update

1. Multimodal and Multi-contex Data Structures
2. Entropy of Binary Trees
3. Entropy of General d-ary Trees
4. Entropy of Graphs
   - Structural Entropy – Unlabeled Graphs
   - Erdős–Renyí Graphs
   - Preferential Attachment Graphs

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

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

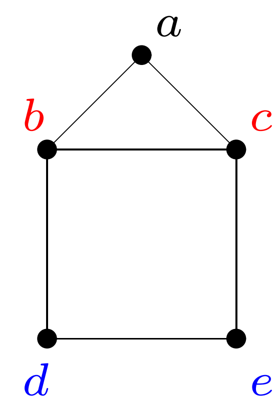
$$H_{\mathcal{S}} = \mathbf{E}[-\log P(S)] = -\sum_{S \in \mathcal{S}} P(S) \log P(S) \quad \text{structural entropy}$$

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 3. *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 4 (Kim, Sudakov, and Vu, 2002)**.** *For Erdös-Rényi graphs and all p satisfying*

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

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 $w > 1$

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

Structural Entropy for Erdös-Rényi Graphs

Theorem 5 (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:* $2^{-\binom{n}{2}(h(p)+\varepsilon)+\log n!} \le P(S) \le 2^{-\binom{n}{2}(h(p)-\varepsilon)+\log n!}$.

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 $\mathcal{G}(n, p)$.

Preferential Attachment Graphs

For an integer parameter m define graph $\mathcal{G}_m(n)$ with vertex set $[n] = \{1, 2, \ldots, n\}$ n the following way:

1. Graph $G_1 \sim \mathcal{G}_m(1)$ is a single node with label 1 with m self-edges.

2. To construct $G_{n+1} \sim \mathcal{G}_m(n+1)$ from G_n:
add vertex $n+1$ and make m random choices $v_1, \ldots, v_m$ as follows:

$$P(v_i = w | G_n, v_1, \ldots, v_{i-1}) = \frac{\deg_n(w)}{2mn},$$

where $\deg_n(w)$ is the degree of vertex $w \in [n]$ in the graph G_n.

Preferential Attachment Graphs

For an integer parameter m define graph $\mathcal{G}_m(n)$ with vertex set $[n] = \{1, 2, \ldots, n\}$ n the following way:

1. Graph $G_1 \sim \mathcal{G}_m(1)$ is a single node with label 1 with m self-edges.

2. To construct $G_{n+1} \sim \mathcal{G}_m(n+1)$ from G_n:
add vertex $n+1$ and make m random choices $v_1, \ldots, v_m$ as follows:

$$P(v_i = w | G_n, v_1, \ldots, v_{i-1}) = \frac{\deg_n(w)}{2mn},$$

where $\deg_n(w)$ is the degree of vertex $w \in [n]$ in the graph G_n.

Symmetry and Automorphism:

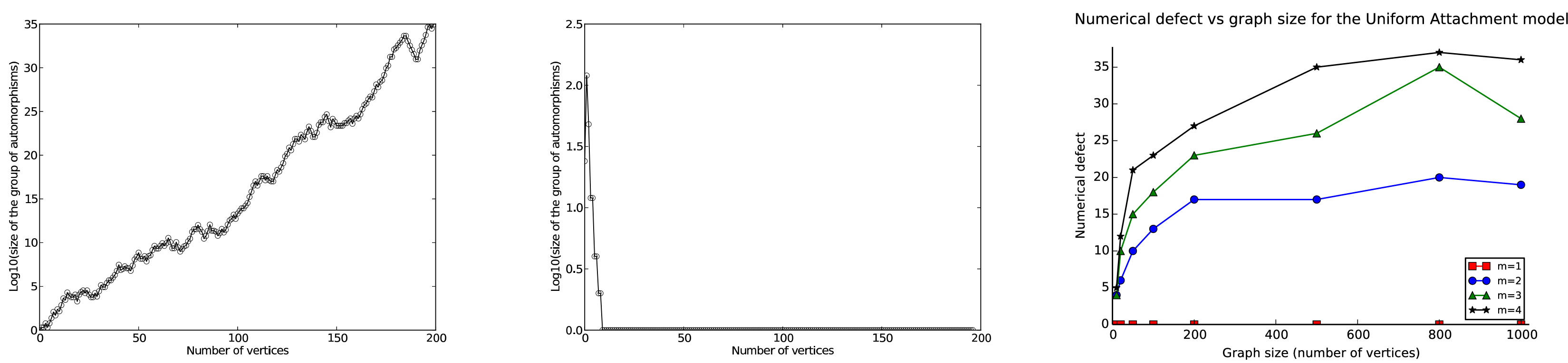


Figure 5: $|\mathrm{Aut}(G)|$: For $m = 1$ (left), $m = 4$ (middle), defect (right).

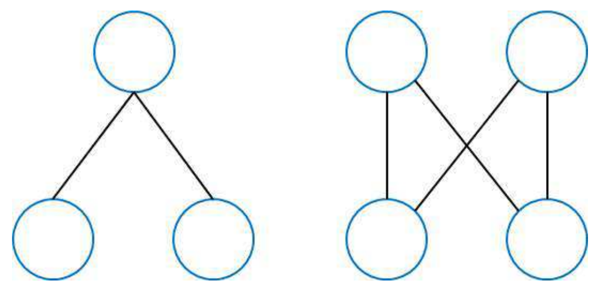
Symmetry of Preferential Attachment Graphs?

Theorem 6 (Janson, Magner, W.S., 2014). *(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 $C > 0$.



That is, for $m = 1, 2$ there are some symmetric subgraphs.

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

$$\Pr[|\mathrm{Aut}(G_n)| > 1] \to 0 \ \ [O(n^{-\varepsilon})],$$

that is, for $m \geq 3$ the graph is asymmetric whp.

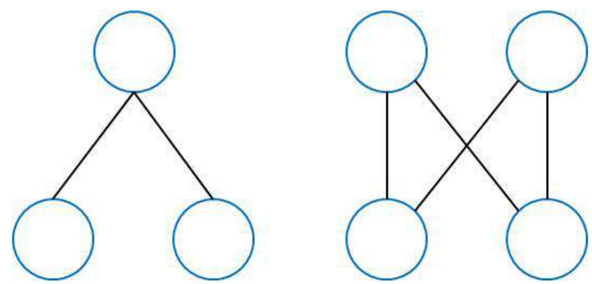
Symmetry of Preferential Attachment Graphs?

Theorem 7 (Janson, Magner, W.S., 2014). *(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 $C > 0$.



That is, for $m = 1, 2$ there are some symmetric subgraphs.

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

$$\Pr[|\mathrm{Aut}(G_n)| > 1] \to 0 \;\; [O(n^{-\varepsilon})],$$

that is, for $m \geq 3$ the graph is asymmetric whp.

Entropy of Preferential Attachment Graphs

Theorem 8 (Luczak, Magner. W.S., 2017). *(Entropy of preferential attachment graphs) Consider $G \sim \mathcal{G}_m(n)$ for fixed $m \geq 1$. We have*

$$H(G) = mn\log n + mn\left(\log 2m - 1 - \log m! - A\right) + o(n),$$

where

$$A = \sum_{d=1}^{\infty} \frac{\log d}{(d+1)(d+2)} \approx 0.868.$$

Theorem 9 (Luczak, Magner, W.S., 2017). *(Structural entropy of preferential attachment graphs) Let $m \geq 3$ be fixed. Consider $G \sim \mathcal{G}_m(n)$. We have*

$$H(S(G)) = (m-1)n\log n + R(n),$$

where $R(n)$ satisfies

$$\left((m+1)\left(\frac{\log 2m}{2} - A\right) + 1\right) n \leq R(n) \leq O(n\log\log n)$$

Remark: Design efficient compression algorithms for labeled and unlabeled Preferential Attachment Graphs matching the above entropy.

That's It

THANK YOU