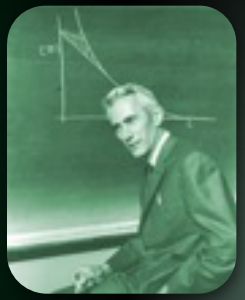
Center for Science of Information



Structural Information: Progress Report

Wojciech Szpankowski

Purdue University

(jointly with A. Grama, A. Magner, and J. Sreedharan)

Bryn Mawr

Howard

MIT

Princeton

Purdue

Stanford

Texas A&M

UC Berkeley

UC San Diego

UIUC

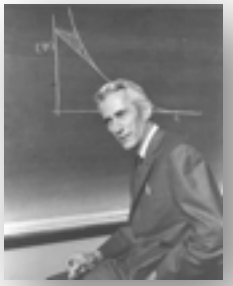
University of Hawaii

Outline

1. **Science of Information**
2. TIMES: Temporal Information Maximally Extracted from Structure
3. Structural Compression
4. TIMES: Recovering Partial Order
5. Experimental Results

   - Synthetic Network
   - Real network
   - Functional Brain Network

What is Science of Information?



- Claude Shannon laid the foundation of information theory, demonstrating that problems of data transmission and compression (i.e., *reliably reproducing data*) can be precisely modeled formulated, and analyzed.

- **SCIENCE OF INFORMATION** builds on Shannon's principles to address key challenges in understanding information that nowadays is not only communicated but also acquired, curated, organized, aggregated, managed, processed, suitably abstracted and represented, analyzed, inferred, valued, secured, and used in various scientific, engineering, and socio-economic processes.

CSoI MISSION: *Advance science and technology through a new quantitative understanding of the representation, communication and processing of information in biological, physical, social and engineering systems.*

Center's Goals

- Extend Information Theory to meet new challenges in

  biology, economics, data & social sciences, and physical distributed systems.

- Understand new aspects of information (embedded) in

  structure, time, space, semantics,

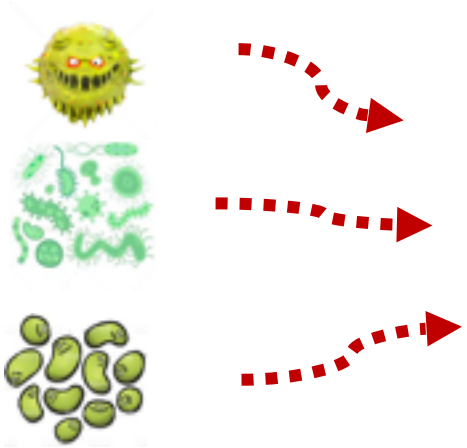
  dynamic information, limited resources, complexity, representation invariant information, and cooperation & dependency.

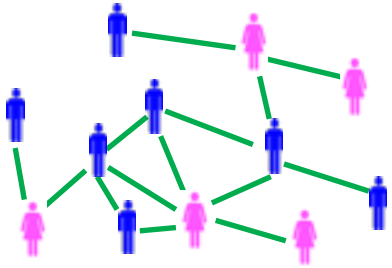
Outline

1. Science of Information
2. **TIMES: Temporal Information Maximally Extracted from Structure**
3. Structural Compression
4. TIMES: Recovering Partial Order
5. Experimental Results

   - Synthetic Network
   - Real network
   - Functional Brain Network

Motivation: Infection Spread





Infection network

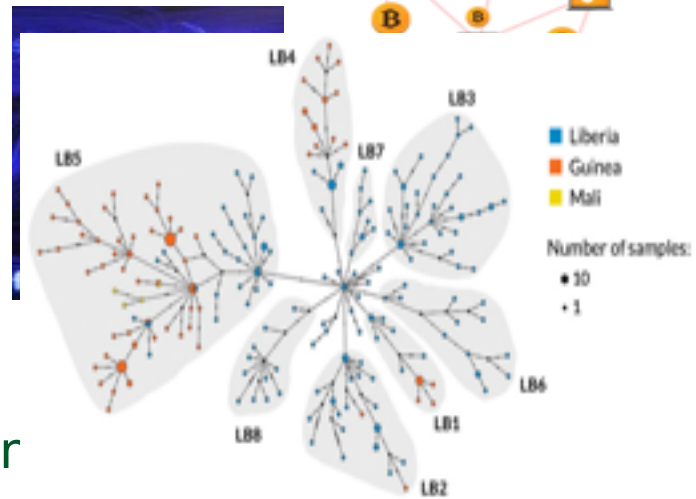
Nodes as patients, edges formed among friends and family.

Only structure info available. Patients admitted not at the order they got infected.

Further Motivation

- Financial transaction networks: flow of capital

- Spread of infectious diseases: origin and initial carriers



- Social networks: spread of information

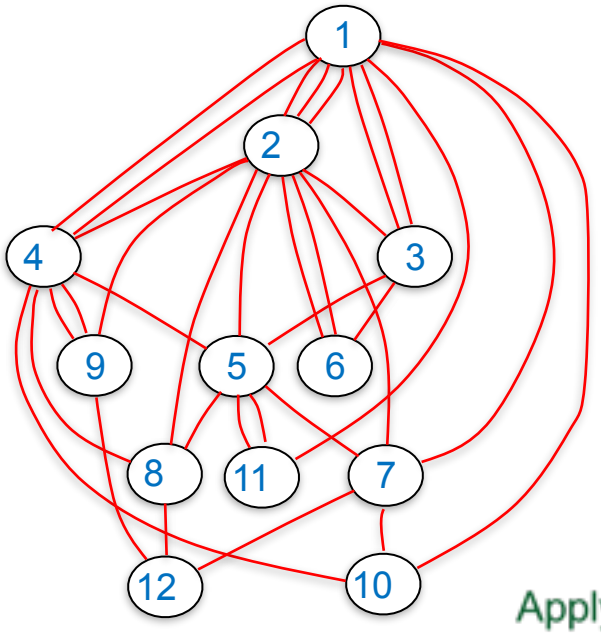
- Network of biochemical reactions:
  (protein-protein interaction network)
  Study of the phylogenetic tree
  Cancer proteins tend to be ancient proteir
  [Srivastava et al., Nature 2010]

Ebola spread network
[Saey, ScienceNews Dec 2015]

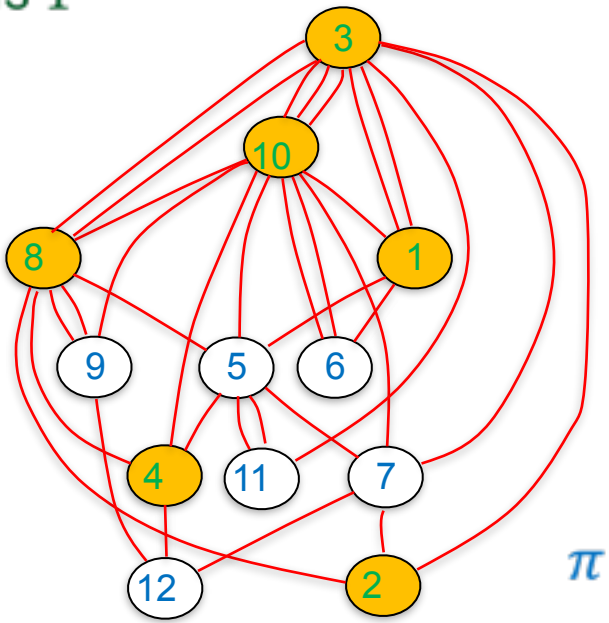
Formulation

Graph G: set of nodes $[n] = \{1, \ldots, n\}$

youngest node label is $n$oldest node label is 1



G



$\pi(G)$

Apply a random permutation π

from S_n. Observed graph is $G' = \pi(G)$

Symmetric group on n letters

Minimax Risk

A node age recovery problem is a tuple $(\mathcal{G}_n, \mathfrak{A}_n, d)$

- $\mathcal{G}_n$: Random graph model
- $\mathfrak{A}_n$: Random adversary function
- d: Distortion measure between permutations

Node age estimator is a function $\phi : \mathfrak{G}_n \to S_n$, i.e., it tries to recover π^{-1}

Set of graphs on n vertices $\longrightarrow$ Symmetry group on n letters

Minimax risk for a random graph model and a distortion measure

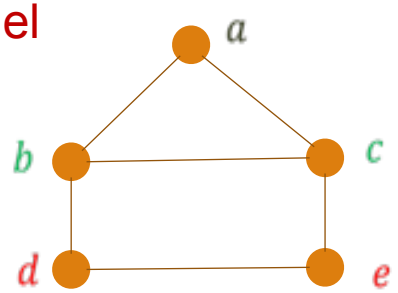
$$R_*(\mathcal{G}_n, d) \stackrel{\Delta}{=} \min_{\phi} \max_{\mathfrak{A}_n} \mathbb{E}[d(\phi(\pi(G)), \pi^{-1})]$$

- d: Distortion measure
- $\pi(G)$: Relabeled graph
- $\mathcal{G}_n$: Random graph model
- ϕ: Estimator
- $\mathfrak{A}_n$: Adversary distribution

Structural Quantities on Graphs

Sets of permutations associated with a random graph model

$\mathrm{Aut}(G)$ Automorphism group of the random graph G distributed according to $\mathcal{G}_n$



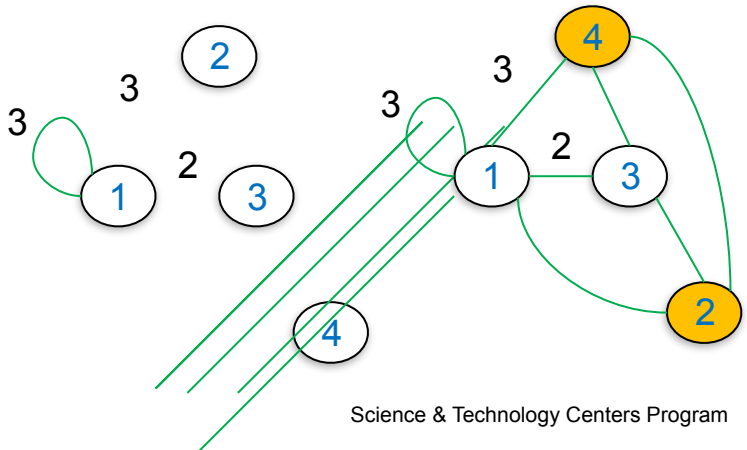
Set of feasible permutations

$\Gamma(G) \subseteq S_n$ Set of permutations σ such that $\sigma(G)$ has positive probability under the distribution $\mathcal{G}_n$

Example of an infeasible permutation:
$\pi = (24) \notin \Gamma(G)$

Set of admissible graphs (unlabeled structures)

$$\mathrm{Adm}(G) \triangleq \{\sigma(G) \ : \ \sigma \in \Gamma(G)\}$$



Lower Bounds on Minimax Risk

Exact recovery $d_e(\sigma_1, \sigma_2) = \mathbb{I}[\sigma_1 \neq \sigma_2]$

Approximate recovery (Kendall Tau distance) $d_a(\sigma_1, \sigma_2) = \sum_{1 \leq i < j \leq n} \mathbb{I}[\sigma_2 \sigma_1^{-1}(i) > \sigma_2 \sigma_1^{-1}(j)]$

Theorem:

For a random graph model $\mathcal{G}_n$ for which any two positive-probability graphs that are isomorphic are equiprobable, the minimax risk for distortion measures d_e and d_a is

Set of admissible graphs

$$R_*(\mathcal{G}_n, d_e) \geq \frac{\mathbb{E}[\log |\mathrm{Aut}(G)|] + \mathbb{E}[\log |\mathrm{Adm}(G)|] - 1}{\log n!}$$

$$= \frac{\mathbb{E}[\log |\Gamma(G)|] - 1}{\log n!}.$$

Set of feasible permutations

Erdős–Rényi & Preferential Attachment Models

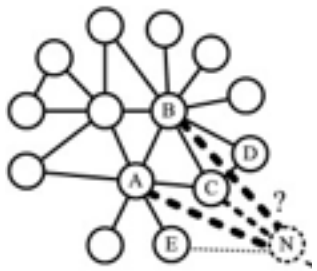
Erdős–Rényi model $G(n,p)$

Each pair of nodes receives an edge independently, with probability p

Preferential Attachment model $\mathcal{PA}(n,m)$

- At $t=1$, G_1 with single vertex (called 1) is created with m self loops
- At $t>1$, vertex t joins and makes m connections to existing nodes in G_{t-1}.
- Each of the m connection choice is independent and satisfies

$$\Pr[t \text{ connects to } k | G_{t-1}] = \frac{\deg_{t-1}(k)}{2m(t-1)}$$



Bad News!

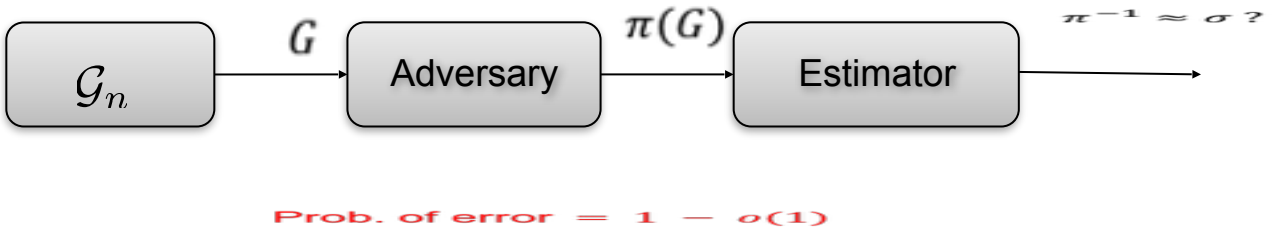
Inapproximability result for Erdős–Rényi and preferential attachment graphs

Let $\mathcal{G}_n$ denote either $G(n,p)$ with $p = p(n) \in (0,1)$ or $\mathcal{PA}(n,m)$, with $m \geq 3$. Then we have

$$R_*(\mathcal{G}_n, d_e) = 1 - o(1).$$

Furthermore, we have, for approximate recovery in the Kendall τ sense,

$$R_*(\mathcal{G}_n, d_a) = \Theta(n^2).$$



Further Bad News!!

ML estimation is not a good approach

$$\mathcal{C}_{\mathrm{ML}}(H) = \arg\max_{\sigma \in S_n} \Pr[G = \sigma^{-1}(H) | \pi(G) = H].$$

Outline

1. Science of Information
2. TIMES: Temporal Information Maximally Extracted from Structure
3. **Structural Compression**
4. TIMES: Recovering Partial Order
5. Experimental Results

- Synthetic Network
- Real network
- Functional Brain Network

Compression of Graphs & Structures

G: Labeled graph
$S(G)$: Unlabeled graph (structure)

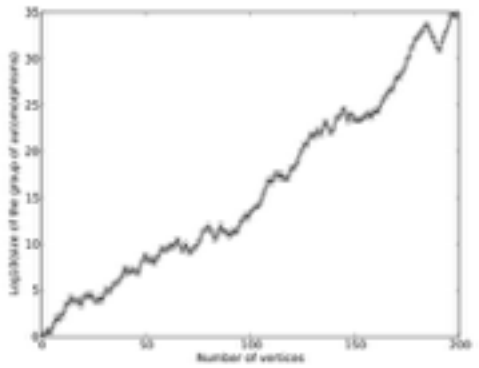
Theorem (Structural entropy for a broad class of graph models)

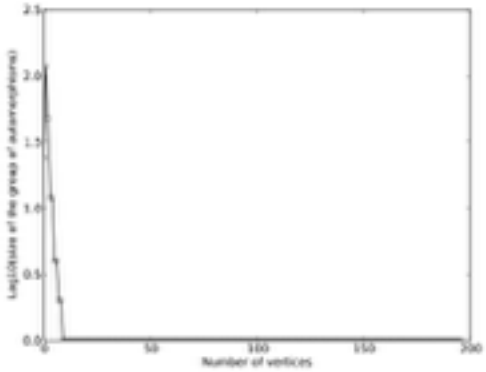
$$H(G) - H(S(G)) = \mathbb{E}[\log|\Gamma(G)|] - \mathbb{E}[\log|\mathrm{Aut}(G)|]$$

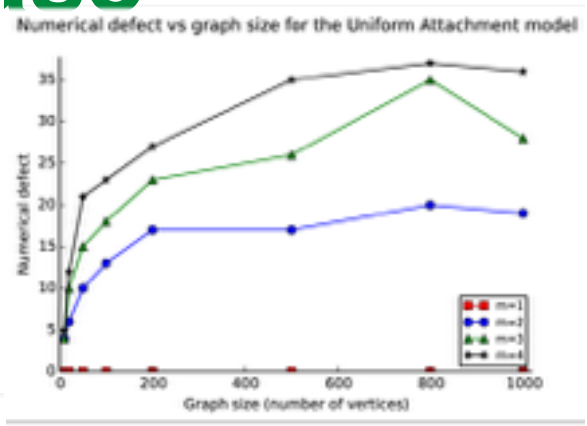
From $\Pr(G|S(G)) = \dfrac{1}{|\mathrm{Adm}(G)|}$ and $|\mathrm{Adm}(G)| = \dfrac{|\Gamma(G)|}{|\mathrm{Aut}(G)|}$

Asymmetry of Preferential attachment case

For m = 1, preferential attachment graph is a tree and is symmetric w.h.p.
For m = 2, preferential attachment graph is symmetric with positive probability



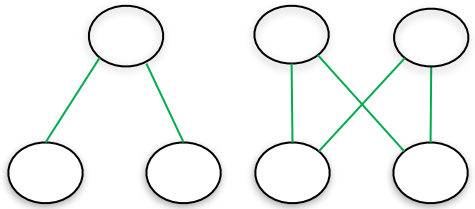




Theorem (Asymmetry for $m \geq 3$)

Consider $G \sim \mathcal{PA}(n, m)$ for $m \geq 3$. Then, for some fixed $\delta > 0$,

$$\Pr[|\mathrm{Aut}(G)| > 1] = O(n^{-\delta})$$



Theorem (Estimate for $|\Gamma(G)|$)

We have, for any $m \geq 1$, if $G \sim \mathcal{PA}(n, m)$,

$$\mathbb{E}[\log |\Gamma(G)|] = n \log n + O(n \log \log n).$$

Structural Entropy for PAG

Theorem: Entropy of preferential attachment graphs

Consider $G \sim \mathcal{PA}(n, m)$ for fixed $m \geq 1$. We have

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

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

Let $m \geq 3$ be fixed. Consider $G \sim \mathcal{PA}(n, m)$. We have

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

Where $R(n)$ satisfies $Cn \leq |R(n)| \leq O(n \log \log n)$ for some nonzero constant $C = C(m)$.

Results of node age recover problem so far are pessimistic

Can we do better ?

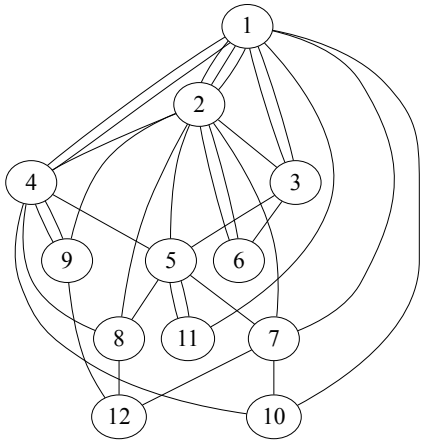
Outline

1. Science of Information
2. TIMES: Temporal Information Maximally Extracted from Structure
3. Structural Compression
4. **TIMES: Recovering Partial Order**
5. Experimental Results

   - Synthetic Network
   - Real network
   - Functional Brain Network

Partial Orders and Binning

Look for partial orders instead of total orders



Precision and Recall

Recall:

Interpret $\pi^{-1}(v)$ as original label of node v of given graph $\pi(G)$

How much we are able to recover?

$$\rho(\phi) = \mathbb{E}\left[\frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} |\{v \in \mathcal{B}_i, w \in \mathcal{B}_j \ : \ \pi^{-1}(v) < \pi^{-1}(w)\}|\right]$$

# of correct pairs

Precision:

How good are the guessed pairs?

$$\theta(\phi) = \mathbb{E}\left[\frac{1}{K(\phi)} \sum_{1 \le i < j \le n} |\{v \in \mathcal{B}_i, w \in \mathcal{B}_j \ : \ \pi^{-1}(v) < \pi^{-1}(w)\}|\right]$$

# of pairs ordered by bins (excluding those inside bins)

$$K(\phi) = \sum_{i,j} |\mathcal{B}_i||\mathcal{B}_j|$$

Density:

$$\delta(\phi) = \frac{\mathbb{E}[K(\phi)]}{\binom{n}{2}}$$

Constrained Optimization Problem

Different approach: phrase as an integer program.

max Precision
Set of partial orders
subject to Density $\geq \varepsilon$

$x_{u,v} : \mathbf{1}\{u <_\phi v\}$ for $u, v \in [n]$
For input DAG H and $\varepsilon > 0$,

max Precision $\Leftrightarrow$ $\max_{\boldsymbol{x}} \dfrac{\sum_{1 \leq u \neq v \leq n} p_{u,v}(H) x_{u,v}}{\sum_{1 \leq u \neq v \leq n} x_{u,v}}$;

Density $\geq \varepsilon \Leftrightarrow$

subject to

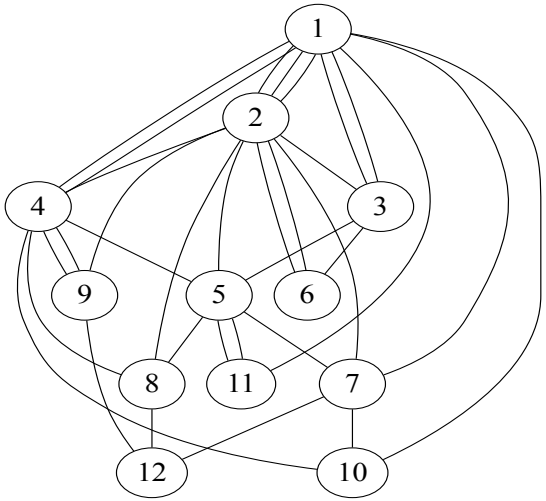
- $x_{u,v} \leq 1 - x_{v,u}$ for all $u \neq v \in [n]$,
- $\sum_{1 \leq u \neq v \leq n} x_{u,v} \geq \epsilon \binom{n}{2}$,
- $x_{u,w} \geq x_{u,v} + x_{v,w} - 1$ for all $u, v, w \in [n]$,
- $x_{u,v} \in \{0, 1\}$ for all $u, v \in [n]$.

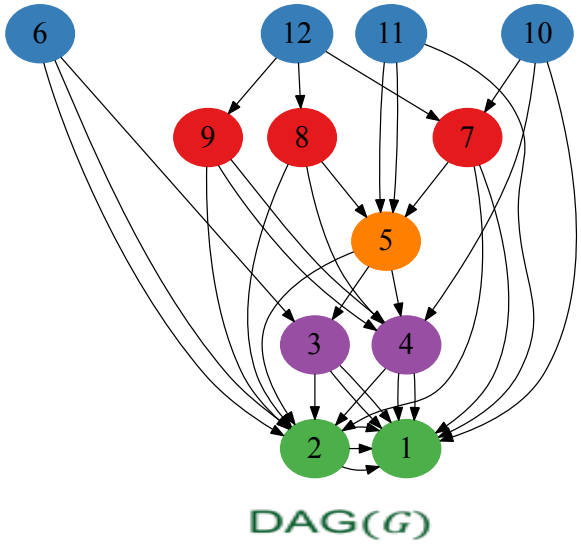
Lemma (coefficients $p_{u,v}(H)$)

We have $p_{u,v}(H) = \Pr[\pi^{-1}(u) < \pi^{-1}(v) | \pi(G) = H] = \dfrac{|\{\sigma : \sigma^{-1} \in \Gamma(H), \sigma^{-1}(u) < \sigma^{-1}(v)\}|}{|\Gamma(H)|}$.

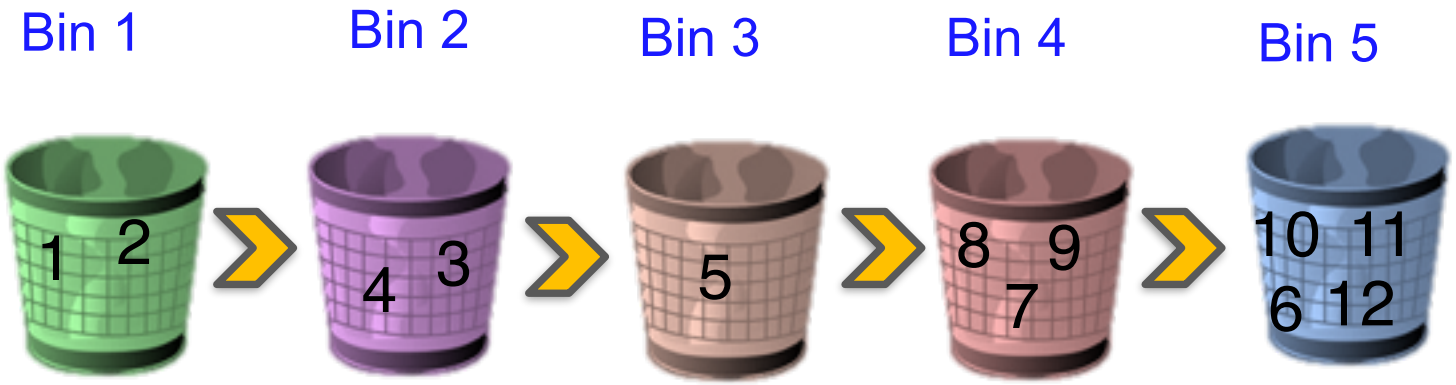
Calculating $p_{u,v}(H)$ is equivalent to counting linear extensions of a partial order
-- #P-complete in general! [Karzanov & Khachiyan, Brightwell & Winkler]
-- Approximate counting in polynomial time (Markov chain algorithm).

Approximating via Peeling algorithm





DAG(G)

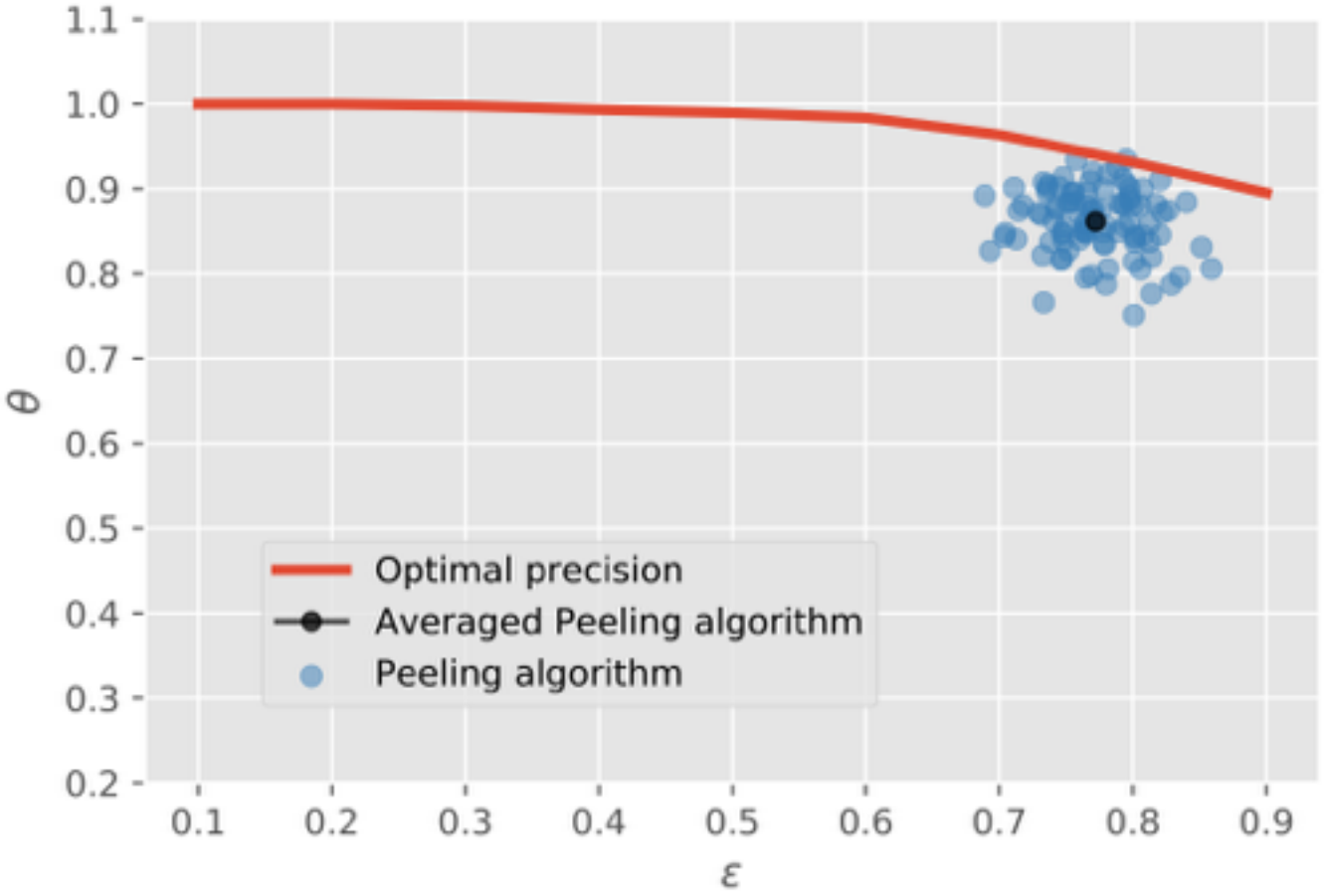


Outline

1. Science of Information
2. TIMES: Temporal Information Maximally Extracted from Structure
3. Structural Compression
4. TIMES: Recovering Partial Order
5. **Experimental Results**
   - Synthetic Network
   - Real network
   - Functional Brain Network

Numerical Results

LP relaxation gives an upper bound.



Theoretical Results

Perfect pair

(u, v) is a perfect pair for a graph H, if for uniformly random permutation π,

$$\Pr[\pi^{-1}(u) < \pi^{-1}(v) | \pi(G) = H] = 1$$

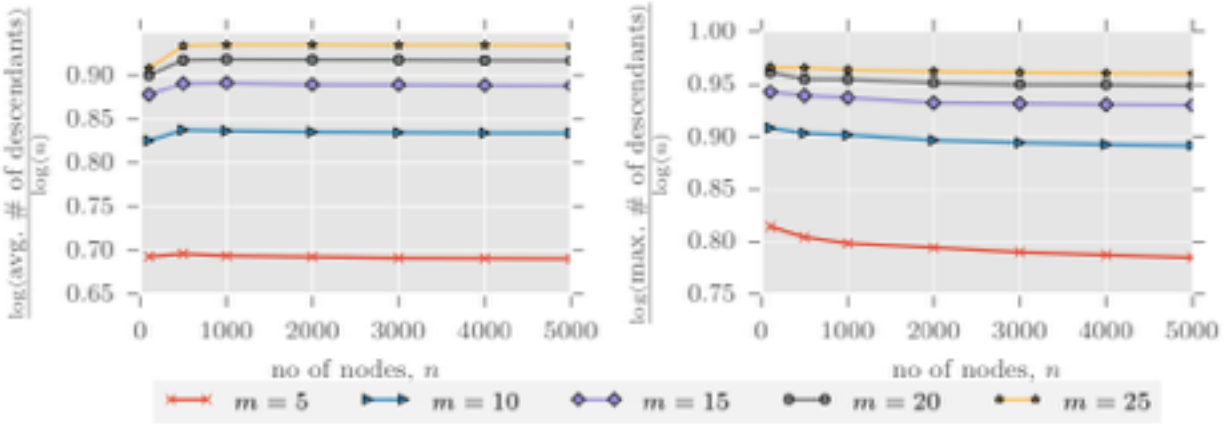
Theoretical Results

for some function $1 > c(m) > 0$.

X_t: Number of perfect pairs in the graph up to time t.

$$\mathbb{E}[X_t] \leq \mathbb{E}[X_{t-1}] + m + \underbrace{O(t^c)}$$

Number of descendants of any given vertex



Experiments: Synthetic Graphs

$n = 5000$

R_{peel}: Ranking with bins given by Peeling algorithm

m	$\theta(R_{\text{peel}})$	$\rho(R_{\text{peel}})$	Bins in R_{peel}
5	0.878	0.758	39.86
10	0.917	0.858	68.77
25	0.958	0.936	140.37
50	0.977	0.967	237.97

How robust is the algorithm?

Technique	$\theta(R_{\text{peel}})$	$\rho(R_{\text{peel}})$	Bins in R_{peel}
$\mathcal{PA}(n, m = 25)$	0.958	0.936	140.37
$\mathcal{PA}(n, M),\ M \sim \text{unif}\{5, 50\}$	0.691	0.683	491.81
$\mathcal{UA}(n, m = 25)$	0.977	0.967	238.23
$\mathcal{UA}(n, M),\ M \sim \text{unif}\{5, 50\}$	0.827	0.823	707.01
Cooper-Frieze (Web graph) model	0.828	0.822	619.25

Uniform Attachment model

$\mathcal{PA} + \mathcal{UA}$+addition of edges between existing nodes

Experiments: Real-World Networks

Citation network (ArXiv High Energy Physics): $|V| = 7.5\text{K}$, $|E| = 116\text{M}$

Simple English Wikipedia: $|V| = 100\text{K}$, $|E| = 1.6\text{B}$

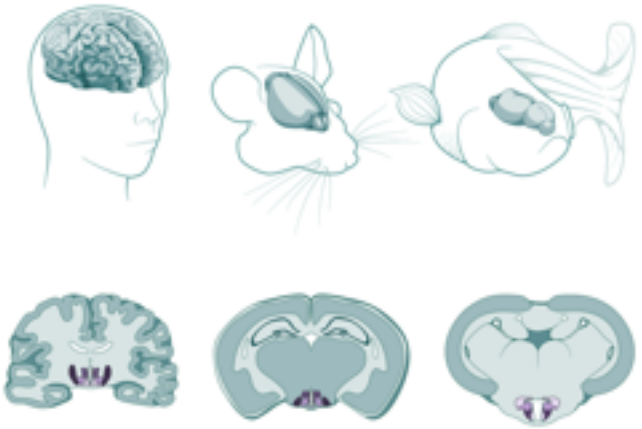
Collaboration network (DBLP Computer Science bibliography):
$|V| = 1.2\text{B}$, $|E| = 5\text{B}$

R_{perf} : Perfect pairs only (nodes with a directed path between them)

	$\theta(R_{\text{perf}})$	$\rho(R_{\text{perf}})$	$\theta(R_{\text{peel}})$	$\rho(R_{\text{peel}})$	Bins in R_{peel}
Citation network	1.0	0.303	0.708	0.681	423
Simple English Wikipedia	0.8968	0.047	0.624	0.548	903
Collaboration network	0.892	0.336	0.785	0.728	18686

Experiments: Brain Networks

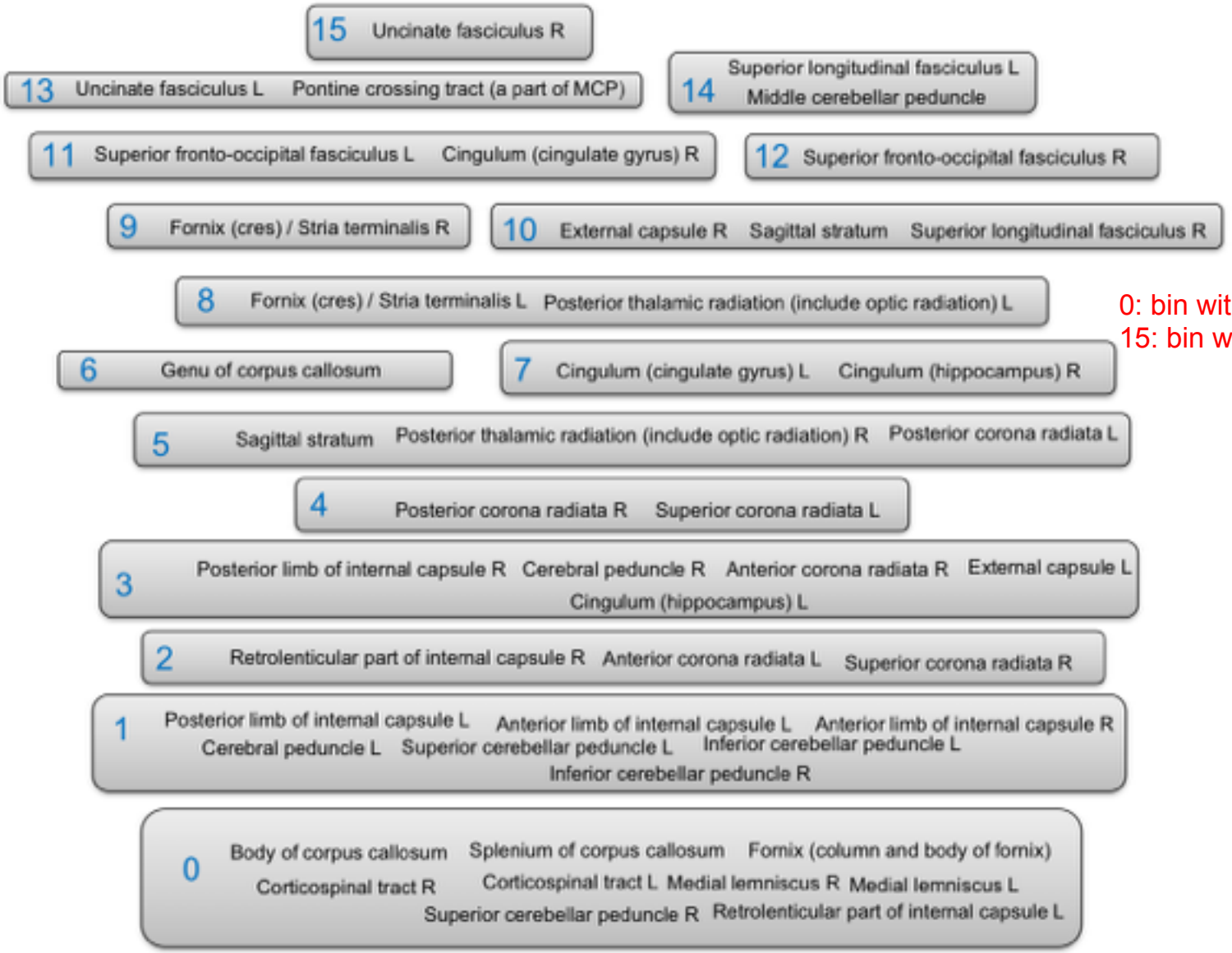
Find age orderings of regions of two different species, based on fMRI images of a same activity.



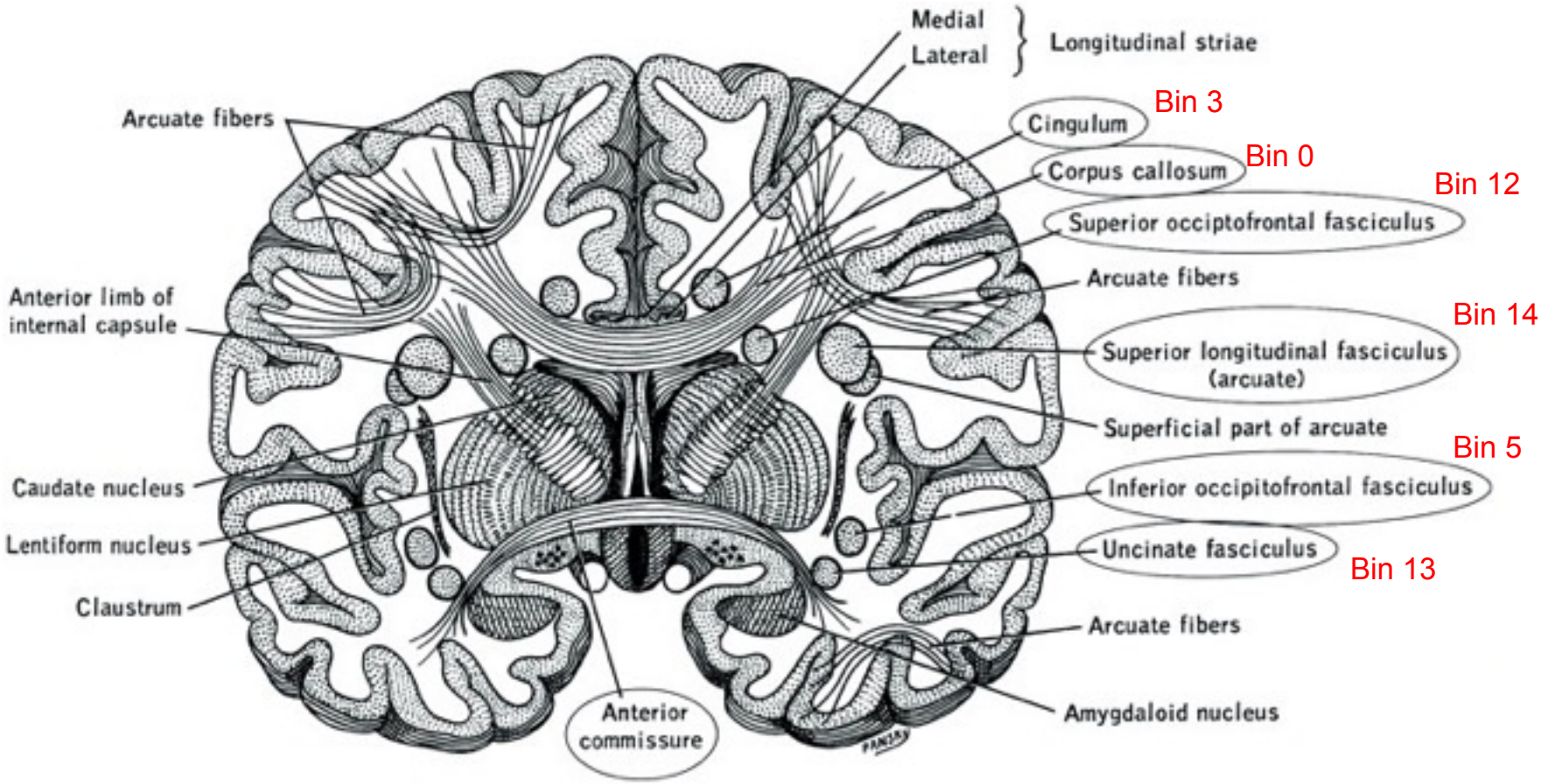
Conjecture: There exists high correlation between the orderings of species evolved from the same genetic parent.

Network formation

- The network has 46 nodes, each of which represents a region in the brain
- An initial network is formed from fMRI images of a human brain in resting state
- Each node here is a voxel and there are 243,648 voxels.
- Each voxel has a time series of data for ~ 350s.
- Pearson correlation coefficient is computed between time series data of each pair of voxels.
- If the correlation > 0.8 we form an edge between the voxels.
- In order to form a network of regions, we make logical OR of the rows and columns in the adjacency matrix of voxel network corresponding to each region.



0: bin with oldest nodes
15: bin with newest nodes


Medial
Lateral
Longitudinal striae
Arcuate fibers
Cingulum
Bin 3
Corpus callosum
Bin 0
Bin 12
Superior occiptofrontal fasciculus
Arcuate fibers
Anterior limb of internal capsule
Bin 14
Superior longitudinal fasciculus (arcuate)
Superficial part of arcuate
Bin 5
Caudate nucleus
Inferior occipitofrontal fasciculus
Lentiform nucleus
Uncinate fasciculus
Bin 13
Claustrum
Arcuate fibers
Amygdaloid nucleus
Anterior commissure