Bayesian Phylogenetic Inference of Stochastic Block Model on Random Graphs

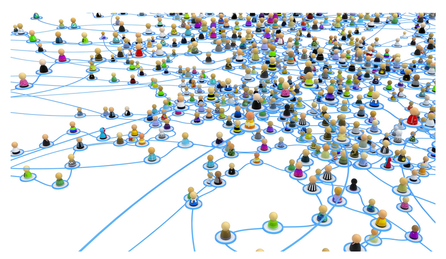
Funded by Center for Undergraduate Research in Mathematics, NSF DMS-1722563

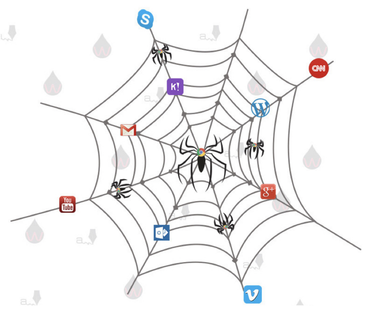
Yanqiu Guo
Joint work with Prof. Wenjian Liu and Ms. Caihong Zheng
Department of Mathematics and Computer Science
Queensborough Community College – The City University of New York

1 Motivations

Networks are ubiquitous:

- Social networks; Biological networks
- Data science: Network sampling; Clustering in graphs

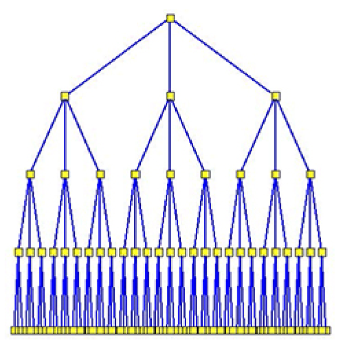


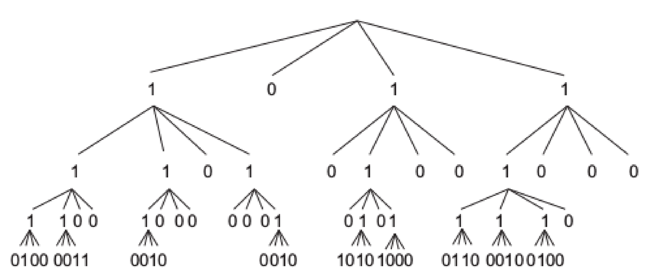


2 Model Setup

Consider the broadcasting process as a discrete, irreducible, aperiodic, and reversible Markov chain.

- $\mathbb{T} = (\mathbb{V}, \mathbb{E}, \rho)$ is a tree with nodes $\mathbb{V}$, edges $\mathbb{E}$ and root $\rho \in \mathbb{V}$
- d-ary tree is the infinite rooted tree where every vertex has exactly d offspring.
- Define a finite characters set $\mathcal{C}$, whose elements are configurations on $\mathbb{T}$, denoted by σ.
- The state of the root ρ, denoted by σ_ρ, is chosen according to an initial distribution π on $\mathcal{C}$.
- Denote by Probability transition matrix $\mathbf{P} = (p_{ij})$ as the noisy communication channel on each edge.
- Let $\sigma(n)$ denote the spins at distance n from the root and let $\sigma^i(n)$ denote $\sigma(n)$ conditioned on $\sigma_\rho = i$.





RECONSTRUCTION: Does this configuration contain a non-vanishing information transmitted by the root, as n goes to ∞?

Definition 1. *The reconstruction problem for the infinite tree $\mathbb{T}$ is solvable if for some $i, j \in \mathcal{C}$,*

$$\limsup_{n\to\infty} d_{TV}(\sigma^i(n), \sigma^j(n)) > 0$$

where d_{TV} is the total variation distance, i.e.

$$d_{TV}(\sigma^i(n), \sigma^j(n)) = \sup_A \left|\mathbf{P}(\sigma(n) = A \mid \sigma_\rho = i) - \mathbf{P}(\sigma(n) = A \mid \sigma_\rho = j)\right|$$

When the $\limsup$ is 0, the model has non-reconstruction *on $\mathbb{T}$.*

3 Background

- The second largest eigenvalue in absolute value of the transition matrix $\mathbf{P}$, say, λ, plays the crucial role in reconstruction problems.
- $d|\lambda|^2 > 1$ (Kesten-Stigum bound): The reconstruction problem is solvable. (Kesten and Stigum, 1966)
- For larger noise $d|\lambda|^2 < 1$: reconstruction depends on the channel.
- The binary symmetric channel: the reconstruction problem is solvable if and only if $d\lambda^2 > 1$ (Bleher et al, 1995).
- The binary-asymmetric model with sufficiently large asymmetry: Mossel (2004) showed that the Kesten-Stigum bound is **NOT** the bound for reconstruction.
- The first exact reconstruction threshold in roughly a decade was obtained by Borgs et al (2006) for the asymmetric Ising channel, i.e.

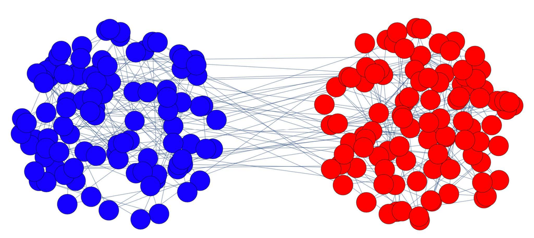
$$\mathbf{P} = \frac{1}{2}\left[\begin{pmatrix} 1+\theta & 1-\theta \\ 1-\theta & 1+\theta \end{pmatrix} + \Delta \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}\right]$$

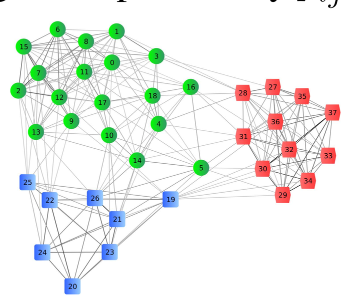
- Liu and Ning (2018) improved the result: when $\Delta^2 > (1-\theta)^2/3$, for every d the Kesten-Stigum bound is not tight. In other words, the reconstruction problem is solvable for some θ even if $d\theta^2 < 1$; when $\Delta^2 < (1-\theta)^2/3$, there exists a $D = D(\pi) > 0$ such that for $d > D$ the Kesten-Stigum bound is sharp.

4 Reconstruction of Stochastic Block Models

4.1 Stochastic Block Models

- The number n of vertices;
- a partition of the vertex set $\{1, \dots, n\}$ into disjoint subsets $C_1, \dots, C_r$, called communities; a symmetric $r \times r$ matrix $\mathbf{P}$ of edge probabilities.
- The edge set is then sampled at random as follows: any two vertices $u \in C_i$ and $v \in C_j$ are connected by an edge with probability p_{ij}.





Different In-block and Out-block Mutations

- Characters set: $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$, consisting of two categories $\mathcal{C}_1 = \{1, \dots, q\}$ and $\mathcal{C}_2 = \{q+1, \dots, 2q\}$.
- The state of the root variable ρ is chosen according to the uniform distribution on $\mathcal{C}$.
- The information flow in the tree according to a general symmetric $2q \times 2q$ probability transition matrix $\mathbf{P}$ with different in-community and out-community transition probabilities, defined as follows:

$$\mathbf{P} = \left(\begin{array}{cccc|cccc} p_0 & p_1 & \cdots & p_1 & p_2 & \cdots & \cdots & p_2 \\ p_1 & p_0 & \cdots & p_1 & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \ddots & \vdots \\ p_1 & p_1 & \cdots & p_0 & p_2 & \cdots & \cdots & p_2 \\ \hline p_2 & \cdots & \cdots & p_2 & \overline{p}_0 & \overline{p}_1 & \cdots & \overline{p}_1 \\ \vdots & \ddots & & \vdots & \overline{p}_1 & \overline{p}_0 & \cdots & \overline{p}_1 \\ \vdots & & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ p_2 & \cdots & \cdots & p_2 & \overline{p}_1 & \overline{p}_1 & \cdots & \overline{p}_0 \end{array}\right)_{2q\times 2q}$$

The eigenvalues of $\mathbf{P}$ are 1 and

$$\lambda_1 = p_0 - p_1, \quad \lambda_2 = p_0 + (q-1)p_1 - qp_2, \quad \lambda_3 = \overline{p}_0 - \overline{p}_1.$$

Kimura 1980 Model

Specially if set $p_0 = \overline{p}_0$, the preceding model becomes K80 DNA sequence evolution model, which has two mutation classes with q states in each class and distinguishes between transitions and transversions.

Theorem 4.1 (Liu et al, 2018). *When $q \geq 4$, for every d the Kesten-Stigum bound is not tight.*

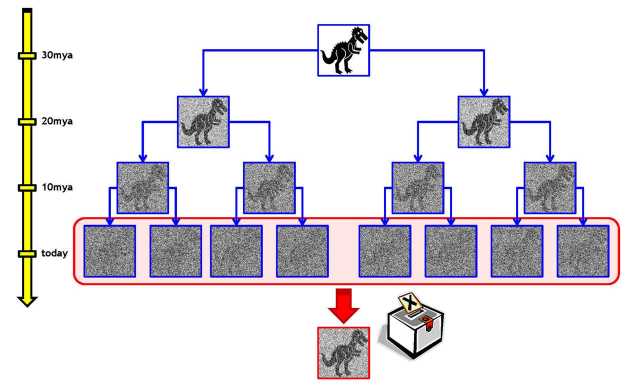
4.2 Research Questions

The corresponding information reconstruction problem in molecular phylogenetics will be explored, by means of the refined analysis of moment recursion on a weighted version of the magnetization, concentration investigation and in-depth investigation on the resulting nonlinear second order dynamical system. Our purpose is to figure out

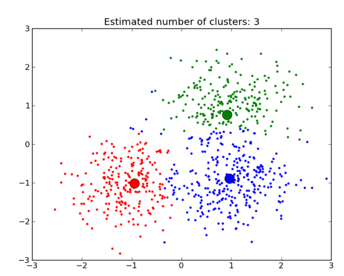
- Under what conditions of $p_0, p_1, \overline{p}_0, \overline{p}_1, p_2$ is the Keston-Stigum reconstruction bound tight, i.e. the reconstruction is unsolvable when $d\lambda_2 < 1$?
- If Keston-Stigum bound is not sharp, then we are interested in figuring out the new reconstruction threshold.

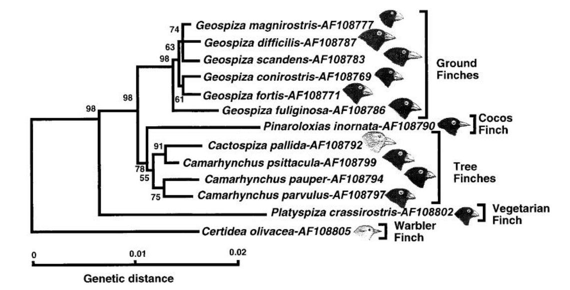
4.3 Application

Unsupervised Learning. Classification problem in unsupervised learning setting using deep generative hierarchical network.



Clustering Problem. Clustering problem in unsupervised learning setting using the stochastic block model.
Phylogenetic Reconstruction. Construct the ancestry tree of a collection of species, given the information of present species.





5 Method and Materials

Let $u_1, \dots, u_d$ be the children of ρ and $\mathbb{T}_v$ be the subtree of descendants of $v \in \mathbb{T}$. Furthermore, if we set $d(\cdot,\cdot)$ as the graph-metric distance on $\mathbb{T}$, denote the nth level of the tree by $L_n = \{v \in \mathbb{V} : d(\rho, v) = n\}$ and then let $\sigma_j(n)$ denote the spins on $L_n \cap \mathbb{T}_{u_j}$. For a configuration A on L_n define the posterior function

$$f_n(i, A) = \mathbf{P}(\sigma_\rho = i \mid \sigma(n) = A) = \mathbf{P}(\sigma_{u_j} = i \mid \sigma_j(n+1) = A).$$

We will research the asymptotic behavior of the objective quantities:

$$x_n = \mathbf{E}\left(f_n(1, \sigma^1(n)) - \frac{1}{2q}\right); \quad \overline{x}_n = \mathbf{E}\left(f_n(q+1, \sigma^{q+1}(n)) - \frac{1}{2q}\right).$$

If the reconstruction problem is solvable, then $\sigma(n)$ contains significant information on the root variable.

Theorem 5.1. *The non-reconstruction is equivalent to*

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} \overline{x}_n = 0.$$

Distributional Recursion

- Key Idea: analyze the recursive relation between $x_n, \overline{x}_n$ and $x_{n+1}, \overline{x}_{n+1}$ by Markov random field property.
- Establish the distributional recursion and moment recursion.
- Display the interactions between spins become very weak if they are sufficiently far away from each other.
- The traditional method is to analyze the the stability of fixed points of the preceding dynamical system of x_n and $\overline{x}_n$.

6 Extensions and Further Discussion

Our technology would be generalized to handle the probabilistic examination of the underlying microscopic systems with large populations on general random graphs.