Lecture Note on Network Statistics

By scribing order: Sayan Das, Ayoushman Bhattacharya, Yi Luo, Hangcen Zou, Yi-Hsuan Shih, Vinh Pham, Giacomo Vedovati, **Anthony Hong**, Samuel Naranjo Rincon, Sidney Nwakanma, Aaron Luo, Adrian Cao, Wei Li, Jingtao Shang, Shourjo Chakraborty, Ty Easley, Haoyu Yin, Zongxi Yu, Bahram Yaghooti, Giacomo Vedovati, Tianwei Zhou, Zhichen Xu, Zongxi Yu, Tong Li [1]

February 17, 2024

[1]Thanks to Professor Lunde for his course Math 586 Network Statistics

Contents

1 Course information **0-7**

2 Notes **1-9**

1.1 Introduction . . . 1-9

1.2 Basics of Graph Theory . . . 1-9

2.3 Representation of a Graph . . . 2-11

2.4 Subgraphs . . . 2-12

3.5 Counting Walks Using the Adjacency Matrix . . . 3-15

3.6 Spectral Graph Theory . . . 3-16

3.6.1 Eigendecomposition . . . 3-16

3.6.2 Single Value Decomposition . . . 3-17

3.6.3 Variational Representations for Eigenvalues . . . 3-17

3.6.4 The Maximum Eigenvalue of A . . . 4-18

4.7 Laplacian Matrix . . . 4-18

4.8 Normalized Matrices . . . 4-20

5.9 Normalized Matrices (Continued) . . . 5-21

5.10 Norms on Matrices . . . 5-22

6.11 Norm on Matrices (Continued) . . . 6-24

6.12 Spectral Perturbation . . . 6-24

6.12.1 Eigenvalue Perturbation . . . 6-24

6.12.2 Eigenvector perturbation . . . 6-25

7.13 Sparsity . . . 7-26

7.14 Degree distribution . . . 7-26

7.15 Transitivity . . . 8-27

8.16 Global Structure of Real-World Networks . . . 8-28

8.17 Local Structure of Networks . . . 8-28
8.17.1 Pagerank Centrality . . . 8-29
8.17.2 Assortativity Coefficient . . . 8-29
9.18 Last Part of Local Structures of Networks . . . 9-30
9.18.1 Measures of Node Importance . . . 9-30
9.18.2 Measure of Similarity . . . 9-30
9.19 Introduction . . . 9-31
9.20 Network Sampling Schemes . . . 9-31
9.21 Sampling Bias . . . 9-31
10.22Horvitz-Thompson Estimator (cont'd) . . . 10-32
10.23Stochastic Process: Markov chain on discrete space . . . 10-33
11.24Introduction . . . 11-36
11.25Markov Properties: . . . 11-36
12.26Introduction . . . 12-38
12.27Lemma: . . . 12-39
13.28Introduction . . . 13-40
14.29Random Walks on graph Review . . . 14-44
14.29.1 Matrix representations . . . 14-44
14.29.2 Total Variaton norm . . . 14-44
14.29.3 Conductance of a cut . . . 14-44
14.30Metropolis-Hasting Algorithm . . . 15-45
15.31Random Graph models . . . 15-45
16.32Exponential Random Graph Models . . . 16-46
17.33Review . . . 17-47
17.33.1 Exponential families . . . 17-47
17.33.2 Canonical form of Exp family . . . 17-47
17.34Exponential Random Graph Modes (ERGMs) . . . 17-47
17.34.1 MGD of Exponential Families . . . 17-48
17.35One alternative method for estimation: MCMCMLE . . . 18-49
18.36Review . . . 18-49
18.37Problem with fitting ERGM . . . 18-50
18.38Pros and Cons of ERGM's . . . 18-50

19.39Block Models 19-51
19.40Stochastic Block Models 19-51
20.41Introduction 20-52
20.42Community Detection algorithms 21-53
21.43Review (Community Detection in SBMs) 21-54
21.44Spectral Clustering (an example) 21-54
21.45Eigenstructure of SBM/Degree-Corrected SBM 21-56
21.46Review 21-56
21.47Spectral Clustering 21-57
21.47.1 Algorithm 21-57
21.47.2 Example 21-57
21.48Eigenstructure of SBM 22-58
22.49Review 22-58
22.50Lemma 22-59
22.51Two Spectral Procedures for DCSBM 23-60
22.52Likelihood-based Methods 23-60
23.53Review 23-60
23.53.1 SBM(α, β) 23-60
23.53.2 Degree-corrected SBM 23-61
23.53.3 Community detection for degree-corrected SBM's 23-61
23.54Likelihood-Based Methods 23-61
23.55Modularity 24-62
24.56Review 24-63
24.57Identifiability issues with RDPG 24-63
25.58Review 25-64
25.59Exchangebility 25-65
28.60Review 28-70
28.60.1 Exchangeability 28-70
28.60.2 Aldous-Hoover Theorem 28-70
28.60.3 Subgraph Frequency 28-71
28.61More Notation for Subgraph Frequency 28-71
28.61.1 Injected Homomorphisms and Induced Homomorphisms 28-71

28.61.2 Convergence 28-71
29.62 Review 29-72
29.62.1 Graph limits 29-72
29.63 Cut Norm 29-72
29.63.1 Relationship between cut metric and subgraph frequencies 30-73
29.64 Szemeredi Regularity lemma 30-73
30.65 Review 30-74
30.66 Theory of Graph Limit 30-74
31.67 Review 31-76
31.68 Deviations in Homomorphism Densities 31-76
31.69 Discussion 32-77
32.70 Review 32-77
32.71 Approaches to Sparse Graph Limits 32-78
32.72 Counting Lemma for L^p graphons 32-79
32.73 Exchangeability 32-79

Chapter 1

Course information

Course description

This is an advanced course on the statistical analysis of network data, covering theory, methodol- ogy, and applications. The course will cover basic graph theory, commonly used models for random graphs (e.g. Erdos-Renyi, exponential random graph models, stochastic block models, random dot product graphs, graphons), and canonical statistical inference problems within these frameworks (e.g. hypothesis testing, community detection, construction of prediction intervals). Time per- mitting, additional topics such as dynamic network modeling and causal inference under network interference will also be discussed. This course is intended for PhD students with appropriately strong mathematical background. However, the course is open to anyone interested in the topic.

Prerequisites

At a minimum, students are expected to be familiar with probability and statistics at the level of Math 493/494. Background in linear algebra at the level of Math 309 is also expected. Graduate level coursework in statistics is a plus. Prior exposure to graph theory is not expected. Proficiency in R/Python/MATLAB is expected for homework assignments.

Chapter 2

Notes

MATH 586	Fall 2023
Statistics for Networks	
Lecture 1: Graph Theory Basics	
Lecturer: Robert Lunde	Scribe: Sayan Das

1.1 Introduction

Why Networks?

- Network: "Collection of interconnected things" (Oxford English Dictionary)
- Modern datasets are not only large but complex: useful to model relationships as networks.
- Networks and graphs are not the same, but graph theory is very useful for studying networks.

Why Statistical Network Analysis?

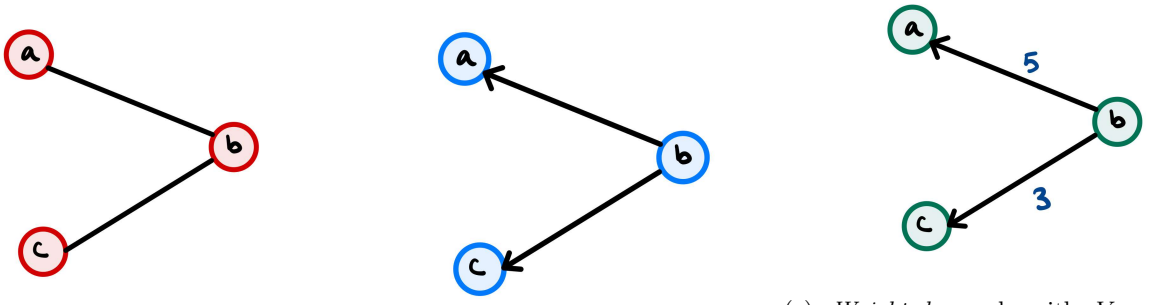
To address scientific/business questions related to network data, we need to be able to do statistics on networks. For example two-sample testing, community detection, and regression with network information.

1.2 Basics of Graph Theory

A *graph* G is defined by the pair (V, E), where V is the vertex set, is the edge set and we write $G = (V, E)$. Note that edges must contain vertices belonging to V.

- For *undirected graphs*, each edge is a set e.g. $\{1, 2\}$.
- For *directed graphs*, each edge is a ordered pair e.g. $(1, 2)$.

- For *weighted graph*, we can define triple (V, E, w), where $w : E \to R$ is a weight function assigning a weight to each edge.



(a) *Undirected* graph with $V = \{a, b, c\}$ and $E = \{\{a, b\}, \{b, c\}\}$

(b) *Directed* graph with $V = \{a, b, c\}$ and $E = \{(b, a), (b, c)\}$

(c) *Weighted* graph with $V = \{a, b, c\}$, $E = \{(b, a), (b, c)\}$ and $W((b, a)) = 5$, $W((b, c)) = 3$

Figure 1.1: Examples of *undirected*, *directed*, and *weighted* graphs.

- Two nodes u and v are *adjacent* if there exists an edge between u and v.
- The *neighborhood* of node u is given by $N(u)$ is the set of neighbors adjacent to node u.

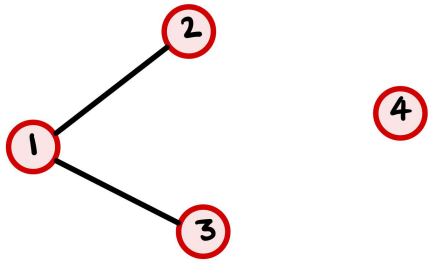


Figure 1.2: Example of a *neighborhood* of a node. Here $N(2) = \{1\}$, $N(1) = \{2, 3\}$, and $N(4) = \{\emptyset\}$. So, the node $\{4\}$ is *isolated*.

- A node u is *isolated* if $|N(u)| = 0$.
- For undirected graphs, the *degree* of node u, often denoted as d_u, is given by $d_u = |N(u)|$.
- For directed graphs, we have two notions of degree:
  - *In-degree* of node u which is the number of edges with endpoint u.
  - *Out-degree* of node u which is the number of edges with starting point u.
- A set is called *independent* if no vertices in the set are adjacent.
- A graph G is *bipartite* if vertices can be divided into two disjoint independent sets V_1 and V_2 such that every edge connects a vertex in V_1 to one in V_2.

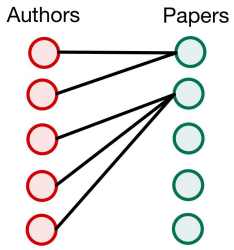


Figure 1.3: Example of *bipartite* graph. Note that, the sets of nodes in the set 'authors' (and also in the set 'papers') are *independent*

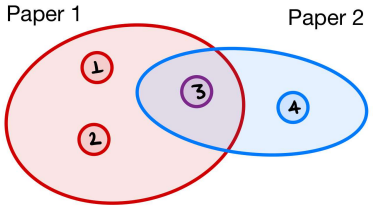


Figure 1.4: Example of *hypergraph* with $V = \{1, 2, 3, 4\}$ and $E = \{\{1, 2, 3\}, \{3, 4\}\}$

- A *hypergraph* $G = (V, E)$ is a graph where edges can have cardinality greater than 2.

Note that Figures 1.2 and 1.3 are alternative ways of expressing the same network. In Figure 1.2, we considered a bipartite graph formulation, where the two disjoint node sets represent authors and papers. In Figure 1.3, we considered a hyperedge setup where papers represent edges (collaborations).

MATH 586 Fall 2023
Statistics for Networks

Lecture 2: Graphs and Subgraphs

Lecturer: Robert Lunde Scribe: Ayoushman Bhattacharya

2.3 Representation of a Graph

A graph $G = (V, E)$ can equivalently be expressed in terms of an *adjacency matrix* $\mathbf{A}$. Consider the vertex set $V = \{1, \ldots, n\}$. Then, the adjacency matrix $\mathbf{A}$ is a $n \times n$ matrix such that

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge from } i \text{ to } j \\ 0, & \text{otherwise} \end{cases}.$$

For undirected graphs, we make do not make distinction between (i, j) and (j, i), that is, $A_{ij} = A_{ji}$; in other words, $\mathbf{A}$ is symmetric. Adjacency matrix for weighted graphs can also be defined similarly.

- Degree of undirected graphs:

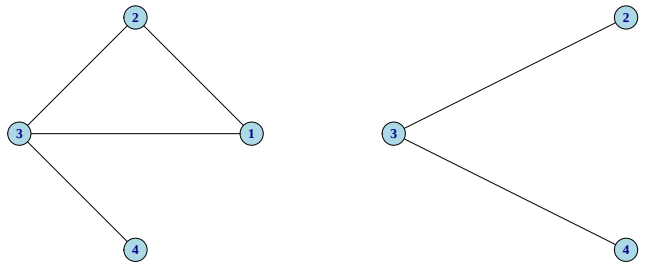
$$d_i = \sum_{j=1}^{n} A_{ij} = \sum_{i=1}^{n} A_{ij}$$

- Typically $A_{ii} = 0$ i.e. no self loops.

2.4 Subgraphs

Given a graph G, we are often interested to subgraphs.

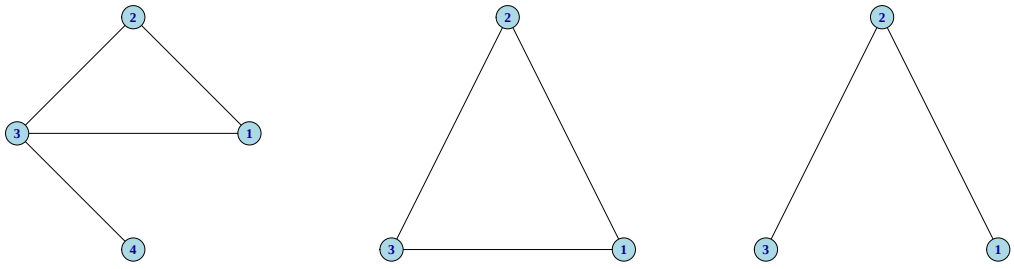
Definition 2.4.1 (Subgraph)**.** A subgraph $H = (V_H, E_H)$ is a graph such that $V_H \subseteq V$ and $E_H \subseteq E$.



(a) Parent graph (b) Subgraph

Figure 2.5: Subgraph of a parent graph

- A subgraph *induced* by the vertex set $U \subseteq V$ has vertex set U and all edges containing vertices U.



(a) Parent graph (b) Induced subgraph on $\{1, 2, 3\}$ (c) Subgraph on $\{1, 2, 3\}$

Figure 2.6: Induced subgraph of a parent graph

- A *clique* is an induced subgraph that is complete.

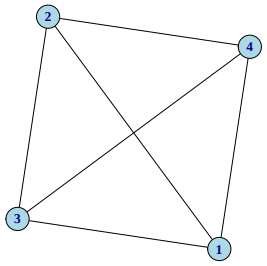


Figure 2.7: Clique with 4 vertices

Definition 2.4.2 (Complete Graph)**.** An undirected graph is complete if every pair of distinct vertices has an edge.

- A *regular* graph is a graph where every vertex has same degree.

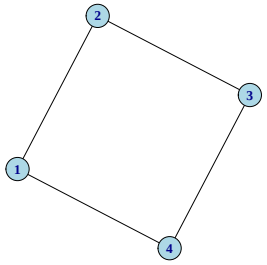
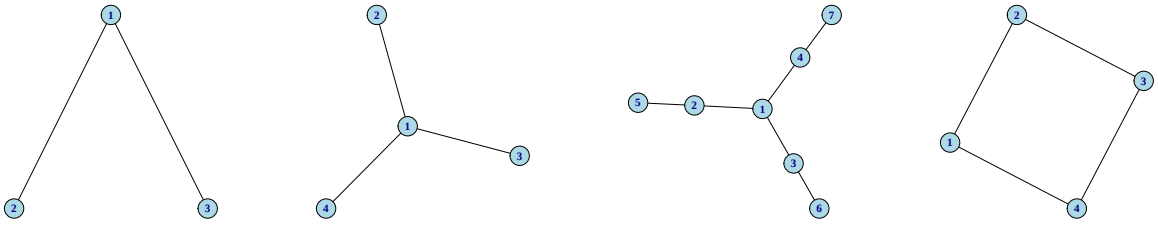


Figure 2.8: A regular graph

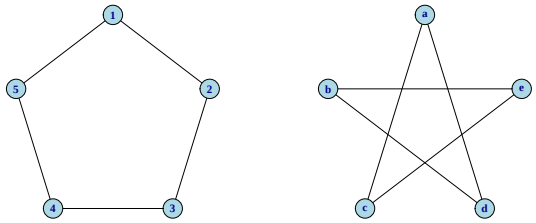
- Here are some examples of other subgraphs that are often used in network literature.



(a) Two-star (b) Three-star (c) Wheel (d) Cycle

Figure 2.9: Different types of subgraphs

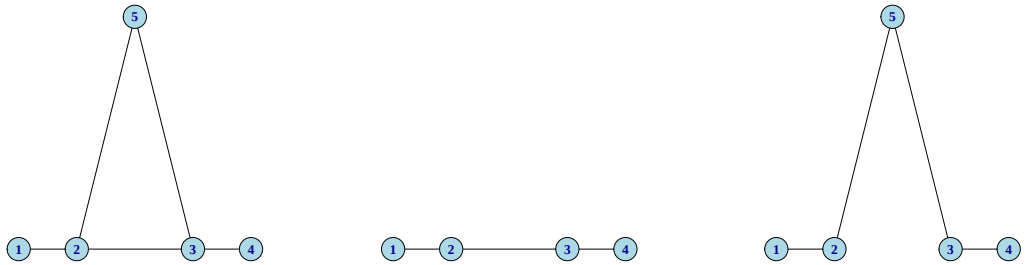
Definition 2.4.3 (Graph Isomorphism)**.** An isomorphism of G and H is a bijection $f : V(G) \mapsto V(H)$ between vertex sets of G and H such that u and v are adjacent if and only if $f(u)$ and $f(v)$ are adjacent.



(a) (b)

Figure 2.10: Graph Isomorphism: $f(a) = 1, f(d) = 2, f(b) = 3, f(e) = 4, f(c) = 5$

- A *walk* is an alternating sequence $\{v_0, e_1, v_1, \ldots, e_{l-1}, v_l\}$ where $e_i = \{v_{i-1}, v_i\}$.



(a) Graph (b) A walk from 1 to 4 (c) A walk from 1 to 4

Figure 2.11: Walk

- *Trails* are walks without repeated edges.
- *Paths* are walks without repeated vertices.
- Cycle is a walk with the same starting and end points.
- A vertex v is *reachable* from u if there exists a walk from u to v.
- A graph is *connected* if every vertex is reachable from another.
  - *Weakly connected:* Undirected version of the graph is connected.
  - *Strongly connected:* Directed graph is connected.
- A (connected) *component* is a maximally connected subgraph.
- Common distance/metric of graphs: *geodesic distance* or shortest path distance, where $d(u, v)$ is given by the length of shortest path between u and v. Distance can be infinity if v is not reachable from u. A metric has to satisfy:

- $d(x, y) = 0 \iff x = y$ (we assume this property for geodesic distance);
- $d(x, y) = d(y, x)$;
  (This is true for undirected graphs, but may not hold for directed graphs)
- $d(x, z) \leq d(x, y) + d(y, z)$.

- *Diameter*: longest path length between vertices. For a graph that is not connected the diameter is infinity.

MATH 586 Fall 2023
Statistics for Networks

Lecture 3: Spectral graph theory I

Lecturer: Robert Lunde Scribe: Yi Luo

3.5 Counting Walks Using the Adjacency Matrix

A^k gives information about the number of walks with length k. For example, for $k = 2$, we see that:

$$A_{ij}^2 = \sum_{k=1}^{n} A_{ik} A_{kj}$$

Every walk of length 2 from i to j must visit an intermediate node before arriving at j. Moreover, when a walk is present, $A_{ik}A_{kj} = 1$ and is 0 otherwise. Since walks with different intermediate nodes are different walks, it is clear that A_{ij}^2 gives us the total number of walks of length 2 from i to j.

Now, for $k = 3$, observe that:

$$(A^3)_{ij} = (A^2 \times A)_{ij} = \begin{bmatrix} \sum_{k=1}^n A_{ik}A_{kj} & \cdots & \sum_{k=1}^n A_{ik}A_{kn} \end{bmatrix} \begin{bmatrix} A_{1j} \\ \vdots \\ A_{nj} \end{bmatrix}.$$

By similar reasoning, every walk of length 3 from i to j must consist of a walk of length 2 and an edge from the last node in the previous walk to node j. Thus, we see that this claim also holds true for $k = 3$. The fact that A^k gives information about the number of walks with length k can be proven for arbitrary k by induction.

The $k = 3$ case is practically important since $\text{trace}(A^3)$ is related to the number of triangles in undirected graphs with no self-loops. In particular, we have

$$\#\text{ triangles} = \text{tr}(A^3)/6.$$

To see this, note that for a given triangle with nodes i, j, k, we could have started the walk from any of the vertices. Moreover, we could have moved clockwise or counterclockwise. Therefore, a given triangle is counted as 6 different walks; by dividing by 6, we adjust for double-counting.

The trace of powers of A can be calculated with help of eigen-decomposition, for example

$$A = VDV^T,\ \text{tr}(A) = \text{tr}(D) = \sum_{i=1}^{n} \lambda_i;$$

$$A^3 = VD^3V^T,\ \text{tr}(A^3) = \text{tr}(D^3) = \sum_{i=1}^{n} \lambda_i^3.$$

3.6 Spectral Graph Theory

We first start by reviewing the notion of eigenvalues and eigenvectors. In what follows, let A be a real, $n \times n$ matrix.

Definition 3.6.1. λ is a (right) eigenvalue with corresponding (right) eigenvector $v \neq 0$ if:

$$Av = \lambda v$$

The above notion corresponds to right multiplication of a matrix A by an appropriate vector v. It is also possible to consider left eigenvalues and eigenvectors. A left eigenvalue/eigenvector pair (λ, v) satisfies:

$$v^T A = \lambda v^T \iff A^T v = \lambda v.$$

Thus, while left and right eigenvalues/eigenvectors may differ in general, they are equivalent when A is symmetric.

The eigenvalues of A are the roots of the characteristic polynomial $p_\lambda(A)$, defined as:

$$p_\lambda(A) = \det(A - \lambda I).$$

The idea is that $Av = \lambda v$ for $v \neq 0$ implies that $(A - \lambda I)$ is singular for appropriate λ and thus the determinant is 0 for these choices of λ. The quantity $p_\lambda(A)$ is a polynomial in λ.

Note that in general, polynomials can have complex roots; thus eigenvalues of real matrices can be imaginary (eigenvectors can have complex numbers as components as well). However, note the following:

Proposition 3.6.2. *Suppose A is a $n \times n$ real, symmetric matrix. Then, all eigenvalues of A are real.*

Since the spectrum of real, symmetric matrices are well-behaved, it is common to consider undirected graphs when one is interested in eigenvalues and eigenvectors.

One may consider a related quantity known as singular values. For a real $m \times n$ matrix A, a singular value and associated left and right singular vectors satisfy:

$$Av = \sigma u \iff A^T u = \sigma v$$

Unlike eigenvalues, singular values are non-negative.

3.6.1 Eigendecomposition

Any real symmetric $n \times n$ matrix admits an eigendecomposition of the form:

$$A = UDU^\top$$

with U orthonormal ($U^\top U = I, UU^\top = I$), and D is a diagonal matrix of the form

$$D = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}, \quad \lambda_1 \geq \cdots \geq \lambda_n.$$

It should be noted that in certain situations, it is more natural to order the elements of D by magnitude instead.

It is also common to express A in outer product form; that is, $A = \sum_{i=1}^n \lambda_i u_i u_i^\top$.

3.6.2 Single Value Decomposition

We now consider SVD, a more general decomposition.

Any $m \times n$ real matrix A permits a factorization of the form:

$$A = \underbrace{U}_{m\times m}\underbrace{D}_{m\times n}\underbrace{V^T}_{n\times n},$$

where D is diagonal and the number of nonzero elements r satisfies $r \leq \min(m, n)$. The nondiagonal elements are the singular values $\sigma_r \leq \ldots \leq \sigma_1$. U and V are both orthonormal; that is, $U^TU = UU^T = I_m$ and $V^TV = VV^T = I_n$.

Using the left and right-singular vectors, A can be written also be written in outer product form as follows:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top.$$

Moreover, using the SVD to express $AA^\top$ (and $A^\top A$), we see that:

$$\begin{aligned} A^\top A &= (UDV^\top)^\top(UDV^\top) = VD^\top U^\top UDV^\top = VD^\top DV^\top \\ AA^\top &= (UDV^\top)(UDV^\top)^\top = UDD^\top U^\top, \end{aligned}$$

We recognize these expressions as eigendecompositions for A^TA and AA^T, respectively. Therefore, we attain the following relation between the singular values of A and the eigenvalues of $AA^\top$ (and $A^\top A$):

$$\sigma_i^2(A) = \lambda_i(AA^\top) = \lambda_i(A^\top A).$$

For symmetric A, if we order the eigenvalues by magnitude, we have the relation:

$$\sigma_i(A) = |\lambda_i(A)|$$

3.6.3 Variational Representations for Eigenvalues

For real symmetric matrices, one can alternatively express the maximum eigenvalue as the solution to a maximization problem:

$$\lambda_{\max} = \max_{\|v\|=1} v^\top Av = \max_{v\neq 0} \frac{v^\top Av}{v^\top v}$$

Of course, the maximum is attained with the corresponding eigenvector v_n. One elementary way of proving this identity is using the method of Lagrange multipliers to solve the constrained optimization problem.

Suppose that $\lambda_n \geq \ldots \geq \lambda_1$. Then, we have the following representation for λ_k:

$$\lambda_k = \max_{\substack{v\neq 0 \\ v_j^\top v=0 \ \forall j:\ k<j\leq n}} \frac{v^TAv}{v^Tv}$$

where $v_1, \ldots v_n$ are the eigenvectors associated with $\lambda_1, \ldots, \lambda_n$. This may also be proved using Lagrange multipliers.

We also have the following result, which states that the eigenvalues are also solutions of an optimization problem over a more general constraint set:

Theorem 3.6.3. *Variational Form/Max-min Form/Courant-Fischer-Weyl Max-min Form:*

Let A be an $n \times n$ real, symmetric matrix with eigenvalues $\lambda_n \geq \cdots \geq \lambda_1$, then

$$\begin{aligned}\lambda_k &= \min_U \left\{ \max_{x \neq 0} \left\{ \frac{x^\top A x}{x^\top x}, x \in U \right\}, \dim(U) = k \right\} \\ &= \max_U \left\{ \min_{x \neq 0} \left\{ \frac{x^\top A x}{x^\top x}, x \in U \right\}, \dim(U) = n - k + 1 \right\}.\end{aligned}$$

3.6.4 The Maximum Eigenvalue of A

Using the properties of eigenvalues derived above, we can now state an important result about the maximum eigenvalue of an adjacency matrix. This result establishes that this eigenvalue is closely related to degrees.

Proposition 3.6.4. *Let A be $n \times n$ adjacency matrix corresponding to an undirected graph. We have:*

$$d_{ave} \leq \lambda_{\max} \leq d_{\max},$$

where d_{ave} denotes the average degree on the graph and $d_{\max}$ denotes the maximum degree on the graph.

Proof. To show $\lambda_{\max} \geq d_{\text{ave}}$,

$$\begin{aligned}\lambda_{\max} &= \max_{v \neq 0} \frac{v^\top A v}{v^\top v} \\ &\geq \frac{\sum_{i=1}^n \sum_{j=1}^n A_{ij}}{\sum_{i=1}^n 1} \\ &\geq d_{\text{ave}}.\end{aligned}$$

where above we chose the all-ones vector 1 to lower bound the variational form.

To prove the upper bound, consider an eigenvector v of the greatest eigenvalue $\lambda_{\max}$. We can choose a component v_j of v such that j takes maximum value of the components $|v_1|, \ldots, |v_n|$. In what follows let $i \sim j$ denote vertex i adjacent to vertex j. Then,

$$|\lambda_{\max}| = \frac{|\lambda_{\max} v_j|}{|v_j|} = \frac{|\sum_{i=1}^n A_{ji} v_i|}{|v_j|} \leq \sum_{i \sim j} \frac{|v_i|}{|v_j|} \leq d_j \leq d_{max}.$$

The result follows. □

MATH 586 Fall 2023
Statistics for Networks

Lecture 4: Spectral graph theory II

Lecturer: Robert Lunde Scribe: Hangcen Zou

4.7 Laplacian Matrix

We now introduce the Laplacian matrix, which is an important object in spectral graph theory. In this lecture, we assume that the adjacency matrix A corresponds to an undirected graph with no self-loops.

Definition 4.7.1 (Laplacian matrix)**.** The Laplacian matrix L is given by $L = D - A$,where $D = \text{diag}(d_1, \ldots, d_n)$. The entries of L are given by:

$$L_{ij} = \begin{cases} d_i & \text{if } i = j \\ -A_{ij} & \text{if } i \neq j \end{cases}$$

To study the Laplacian matrix, it is helpful to consider an object known as the (oriented) incidence matrix. To construct this oriented incidence matrix, for each $\{i, j\} \in E$, we pick an orientation, where one node is the head, the other is the tail (note that the choice is completely arbitrary). We have the following notion:

Definition 4.7.2 (Incidence matrix)**.** Define the $|V| \times |E|$ incidence matrix B as

$$B_{ij} = \begin{cases} 1 & \text{if } i \text{ is in } j^{th} \text{ edge, } i \text{ is head} \\ -1 & \text{if } i \text{ is in } j^{th} \text{ edge, } i \text{ is tail} \\ 0 & \text{o.w.} \end{cases}$$

Now observe that

$$L = BB^\top.$$

To see this, note that for $i \neq j$, the term $B_{ik}B_{jk}$ is nonzero if and only if nodes i and j belong to the kth edge. Moreover, when it is nonzero, one vertex is the head and the other is the tail. Thus,

$$(BB^\top)_{ij} = \sum_{k=1}^{|E|} B_{ik}B_{jk} = -A_{ij}.$$

For $i = j$, each possible edge involving node i is counted once. Therefore,

$$(BB^\top)_{ii} = \sum_{k=1}^{|E|} B_{ik}^2 = d_i$$

Reall that a $n \times n$ matrix A is positive semidefinite if $x^\top Ax \geq 0 \; \forall \; x$. We claim the following:

Proposition 4.7.3. *L is positive semidefinite.*

Proof. By using the representation $L = BB^T$, we have:

$$\begin{aligned} x^\top(BB^\top)x &= (B^\top x)^\top(B^\top x) \\ &= \sum_{(i,j)\in E} (x_i - x_j)^2 \geq 0 \end{aligned}$$

□

Recall that positive semi-definiteness implies that all eigenvalues of the matrix are non-negative. It will turn out that the eigenvalue zero of the Laplacian gives us substantial information about the connectivity of the graph.

Before we explore these properties, note the following variational characterization of the minimum eigenvalue.

Proposition 4.7.4. *Suppose that A is a real symmetric $n \times n$ matrix. Then,*

$$\lambda_{min} = \min_{\|x\|=1} x^T Ax = \min_{x \neq 0} \frac{x^T Ax}{x^T x}$$

The following proposition establishes that L has at least one 0 eigenvector and provides a corresponding eigenvector.

Proposition 4.7.5. *The vector* $1 = (1, \ldots, 1)$ *is always an eigenvector of* L*, with eigenvalue* 0*.*

Proof. Note that for 1, we have that:

$$1^T L 1 = \sum_{(i,j)\in E} (1-1)^2 = 0$$

Since $\vec{1}$ attains the minimum value of $x^T Lx$, it follows that it must be an eigenvector of the eigenvalue 0. □

We now state the following result regarding the multiplicity of the zero eigenvalue.

Proposition 4.7.6. *The multiplicity of the eigenvalue* 0 *for* L *gives the number of connected components in the graph.*

Proof. **Step 1:** Show # zero eigenvalues $\geq$ # connected components in graph.
For each connected component, consider a vector that is constant on the connected component and is 0 for all other entries.

It is clear that these vectors are orthogonal. Moreover, since there are no edges between vertices in different connected components, it is clear that for these vectors:

$$v^T L v = \sum_{(i,j)\in E} (v_i - v_j)^2 = 0$$

Thus, they correspond to zero eigenvalues.

Step 2: # zero eigenvalues $\leq$ # connected components.
We argue by contradiction. Suppose there exists another zero eigenvector v_{k+1}, where $v_{k+1} \neq 0$. Now we claim that v_{k+1} must be constant on connected components.

To see this, suppose that it is not constant on a connected component. Suppose that the entries of the eigenvector corresponding to this component can be grouped into K groups taking distinct values, where $K > 2$ by assumption. Since the component is connected, there must be a path from one group to the other. Thus for this edge, $(v_i - v_j)^2 > 0$ and thus v is not an 0 eigenvector, a contradiction.

Now, suppose v_{k+1} is constant on i^{th} component, then v_{k+1} and v_i are not orthogonal, which is a contradiction. □

4.8 Normalized Matrices

Normalized versions of the adjacency matrix and Laplacian also play an important role in spectral graph theory. We define these notions below.

Definition 4.8.1 (Normalized adjacency matrix)**.** The normalized adjacency matrix $\mathcal{A}$ is given by

$$\mathcal{A} = D^{-1/2} A D^{-1/2}, \quad \text{where } D^{-1/2} = \text{diag}(d_1^{-1/2}, \ldots, d_n^{-1/2})$$

Definition 4.8.2 (Normalized Laplacian)**.** The normalized Laplacian matrix $\mathcal{L}$ is defined as:

$$\mathcal{L} = I - \mathcal{A},$$

where L satisfies:

$$\mathcal{L}_{ij} = \begin{cases} 1 & i = j \\ -\frac{1}{\sqrt{d_i}\sqrt{d_j}} & i \neq j,\ A_{ij} = 1 \\ 0 & o.w. \end{cases}$$

Note that $\mathcal{L}$ can also be expressed as:

$$\mathcal{L} = D^{-1/2} L D^{-1/2}$$

Let α_i denote eigenvalue i of $\mathcal{A}$, λ_i equal eigenvalue i of $\mathcal{L}$, we have

$$1 = \alpha_1 \geq \ldots \geq \alpha_n \geq -1$$
$$0 \leq \lambda_1 \leq \ldots \leq \lambda_n \leq 2$$

We will prove some of these properties next class.

MATH 586 Fall 2023
Statistics for Networks

Lecture 5: Matrix Norms

Lecturer: Robert Lunde Scribe: Hangcen Zou

5.9 Normalized Matrices (Continued)

Let $\mathcal{A}$ denote the normalized adjacency matrix and $\mathcal{L}$ denote the normalized Laplacian matrix, then it can be shown that:

$$\begin{aligned} x^\top \mathcal{L} x &= x^\top (I - \mathcal{A}) x \\ &= \sum_{i \in V} x_i^2 - \sum_{(i,j) \in E} \frac{2 x_i x_j}{\sqrt{d_i}\sqrt{d_j}} \\ &= \sum_{(i,j) \in E} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2 \\ &\geq 0 \end{aligned}$$

Thus, $\mathcal{L}$ is also positive semidefinite. From the above expression, it is clear that $D^{1/2}1$ is an eigenvector corresponding to the eigenvalue 0.

Moreover, we see that for any eigenvector of $\mathcal{L}$ corresponding to a 0 eigenvalue, we see that we can choose

Proposition 5.9.1. *Let α_i denote eigenvalue i of $\mathcal{A}$, and λ_i denote eigenvalue i of $\mathcal{L}$, then we have*

$$1 = \alpha_1 \geq \ldots \geq \alpha_n \geq -1$$
$$0 \leq \lambda_1 \leq \ldots \leq \lambda_n \leq 2$$

Proof. By utilizing the positive semi-definite property, we have

$$\begin{aligned} x^\top \mathcal{L} x &\geq 0 \\ x^\top (I - \mathcal{A}) x &\geq 0 \\ x^\top x - x^\top \mathcal{A} x &\geq 0 \\ x^\top x &\geq x^\top \mathcal{A} x \\ 1 &\geq \frac{x^\top \mathcal{A} x}{x^\top x}. \end{aligned}$$

Similarly,

$$\begin{aligned} x^\top (I + \mathcal{A}) x &\geq 0 \\ x^\top x &\geq -x^\top \mathcal{A} x \\ \frac{x^\top \mathcal{A} x}{x^\top x} &\geq -1 \end{aligned}$$

□

5.10 Norms on Matrices

Definition 5.10.1 (Norm)**.** Norm $\|\cdot\|$ is a function $\mathcal{X} \mapsto \mathbb{R}$ that satisfies

1. $\|x\| \geq 0$
2. $\|x\| = 0 \mathcal{L} eftrightarrow x = 0$
3. $\|\alpha x\| = |\alpha| \cdot \|x\|$, where α is a scalar
4. $\|x + y\| \leq \|x\| + \|y\|$

Definition 5.10.2 (Frobenius norm)**.** Let B be an $m \times n$ matrix, the Frobenius norm of B, denoted $\|B\|_F$, is given by

$$\begin{aligned} \|B\|_F &= \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} B_{ij}^2} \\ &= \sqrt{trace(BB^\top)} \\ &= \sqrt{trace(U \Sigma U^\top)} \qquad \text{by eigendecomposition} \\ &= \sqrt{\sum_{i=1}^{r} \sigma_i^2(B)} \qquad r \text{ represents the rank} \end{aligned}$$

Definition 5.10.3 (Operator norm)**.**

$$\|B\|_{op} = \sup_{\|v\|=1} \|Bv\| = \sup_{v \neq 0} \frac{\|Bv\|}{\|v\|}$$

Proposition 5.10.4. $\|B\|_{op} = \sigma_{max}(B)$

Proof.

$$\|B\|_{op}^2 = \left(\sup_{\|v\|=1} \|Bv\|\right)^2 = \sup_{\|v\|=1} \|Bv\|_2^2$$
$$= \sup_{v\neq 0} \frac{(Bv)^\top Bv}{v^\top v} = \sup_{v\neq 0} \frac{v^\top (B^\top B)v}{v^\top v} = \sigma_{max}^2(B)$$

□

Proposition 5.10.5. $\|B\|_{op} \leq \|B\|_F$

Proposition 5.10.6. $\|Bv\| \leq \|B\|_{op}\|v\|$

Proof. $\|Bv\| \leq \sup_{v\neq 0} \frac{\|Bv\|}{\|v\|}\|v\|$ □

Proposition 5.10.7. $\|AB\|_{op} \leq \|A\|_{op}\|B\|_{op}$

Proof. $\frac{\|A(Bv)\|}{\|Bv\|}\|Bv\| \leq \sup_{Bv\neq 0} \frac{\|A(Bv)\|}{\|Bv\|} \cdot \sup_{\|v\|=1} \|Bv\|$ □

Proposition 5.10.8. $\|AB\|_F \leq \|A\|_F\|B\|_{op}$ *and* $\|AB\|_F \leq \|A\|_{op}\|B\|_F$

Proof.

$$\|AB\|_F^2 = \sum_{i=1}^n \|Ab_i\|^2 \leq \sum_{i=1}^n \|A\|_{op}^2\|b_i\|^2$$
$$= \|A\|_{op}^2 \sum_{i=1}^n \|b_i\|^2 = \|A\|_{op}^2\|B\|_F^2$$

□

Proposition 5.10.9. *Frobenius norm is invariant under orthogonal transformations. Suppose* $U^\top U = I$*, then* $\|UM\|_F = \|M\|_F$

Proof.

$$\|UB\|_F = \sqrt{trace\big((UB)(UB)^\top\big)} = \sqrt{trace\big(UBB^\top U^\top\big)}$$
$$= \sqrt{trace(BB^\top)} = \|B\|_F$$

□

The following are other common matrix norms,

$$\|B\|_\infty = \max_{i,j} |B_{ij}|$$
$$\|B\|_{2\to\infty} = \max_i \|b_i\|_2.$$

For finite dimensional vector spaces, we have equivalentce of norms. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be two norms, then there exist constants c_1, c_2 such that

$$c_1\|x\|_\alpha \leq \|x\|_\beta \leq c_2\|x\|_\alpha \quad \forall x$$

MATH 586 Fall 2023
Statistics for Networks

Lecture 6: Spectral Perturbation

Lecturer: Robert Lunde Scribe: Yi-Hsuan Shih

6.11 Norm on Matrices (Continued)

Proposition 6.11.1. *If $UU^T = I$, then $\|UA\|_{op}^2 = \|A\|_{op}^2$*

Proof.

$$\|UA\|_{op}^2 = \sup_{x \neq 0} \frac{(UAx)^T(UAx)}{x^T x} = \sup_{x \neq 0} \frac{x^T A^T U^T U A x}{x^T x} = \sup_{x \neq 0} \frac{x^T A^T A x}{x^T x} = \|A\|_{op}^2$$

□

6.12 Spectral Perturbation

Suppose our $n \times n$ symmetric matrix can be expressed as

$$A = \underbrace{M}_{signal} + \underbrace{E}_{noise}.$$

Q: Are eigenvalues/eigenvectors of A "close" to M?

With PCA, we are interested in eigenvalues/eigenvectors of

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T, \quad \mathbb{E}[X_i] = 0, \quad X_i \in \mathbb{R}^p,$$
$$\Sigma = \mathbb{E}[XX^T],$$
$$\hat{\Sigma} = \Sigma + E.$$

In fixed dimensions, $\hat{\Sigma} - \Sigma \xrightarrow{P} 0$.

Also true (under different conditions for different norms) when d is growing with n.

6.12.1 Eigenvalue Perturbation

Theorem 6.12.1 (Weyl's inequality)**.** *Suppose A, B are real symmetric matrices with eigenvalues $\lambda_n \geq \cdots \geq \lambda_1$, and $\gamma_n \geq \cdots \geq \gamma_1$, respectively. Then*

$$\max_{1 \leq i \leq n} |\lambda_i - \gamma_i| \leq \|A - B\|_{op}.$$

Example 6.12.2. $A = \hat{\Sigma}$, $B = \Sigma$. Thus

$$\max_{1 \leq i \leq n} |\lambda_i(\hat{\Sigma}) - \lambda_i(\Sigma)| \leq \|\hat{\Sigma} - \Sigma\|_{op}.$$

Theorem 6.12.3 (Hoffman-Wielandt inequality)**.** *Under the same conditions as Weyl's theorem,*

$$\sum_{i=1}^{n}(\lambda_i - \gamma_i)^2 \leq \|A - B\|_F^2.$$

Q: In typical statistical applications, how large is $\|\hat{\Sigma} - \Sigma\|_{op}$?

A: If $\Sigma = I$ and X "light-tailed", then we have

$$\|\hat{\Sigma} - \Sigma\|_{op} \leq \sqrt{\frac{d}{n}} + \frac{d}{n} \quad w.h.p,$$
$$\|\hat{\Sigma} - \Sigma\|_{\infty} \leq \sqrt{\frac{\log d}{n}} \quad w.h.p.$$

6.12.2 Eigenvector perturbation

When are, for example, eigenvectors of $\hat{\Sigma}$ and Σ close?

How can we compare $v_1, \ldots, v_k$ with $\hat{v}_1, \ldots, \hat{v}_k$?

Recall:

$$a^T b = \|a\|\|b\| cos\theta,$$
$$\text{If } \|a\| = \|b\| = 1,\ a^T b = cos\theta.$$

Definition 6.12.4 (Canonical angle)**.** Let E be a matrix with orthonormal columns, F another with orthonormal columns. The first canonical angle is given by

$$\begin{aligned}
\theta_1 &= \cos^{-1}\left(\sup_{\substack{x\in col(E),\, y\in col(F)\\ \|x\|=1,\, \|y\|=1}} x^T y\right) \\
&= \cos^{-1}\left(\sup_{x\neq 0,\, y\neq 0} \frac{(Ex)^T(Fy)}{\|x\|\|y\|}\right) \\
&= \cos^{-1}\left(\sup_{x\neq 0,\, y\neq 0} \frac{x^T(E^T F)y}{\|x\|\|y\|}\right) \\
&= \cos^{-1}\left(\sigma_{max}(E^T F)\right).
\end{aligned}$$

where $\sigma_{max}(A) = \sup_{x\neq 0,\, y\neq 0} \frac{x^T A y}{\|x\|\|y\|}$ is the maximum singular value.

The k^{th} canonical angle is given by

$$\begin{aligned}
\theta_k &= \cos^{-1}\left(\sup_{\substack{x\in col(E),\, y\in col(F)\\ \|x\|=1,\, \|y\|=1\\ x^T x_r = 0,\, y^T y_r = 0,\, \forall 0<r<k}} x^T y\right) \\
&= \cos^{-1}\left(\sigma_k(E^T F)\right).
\end{aligned}$$

Definition 6.12.5. Let $\sin\Theta = diag(\sin\theta_1, \ldots, \sin\theta_k)$. This turns out that $\|\sin\Theta\|_F$ is a metric on $d-$dimensional linear space.

Theorem 6.12.6 (Davis-Kahan $\sin\Theta$ theorem)**.** *Let $\hat{\Sigma}$, $\Sigma \in \mathbb{R}^{p\times p}$ be symmetric with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$, and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$, receptively. Fix $1 \leq r \leq s \leq p$, and let $d = s - r + 1$ and $V = (v_r, v_{r+1}, \ldots, v_s)$, $\hat{V} = (\hat{v}_r, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ satisfying $\Sigma v_j = \lambda_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\lambda}_j \hat{v}_j$. If $\delta = \inf\{|\hat{\lambda} - \lambda| : \lambda \in [\lambda_s, \lambda_r], \hat{\lambda} \in (-\infty, \hat{\lambda}_{s-1}]\bigcup[\hat{\lambda}_{r+1}, \infty)\} > 0$, then*

$$\|\sin\Theta|_F \leq \frac{\|\hat{\Sigma} - \Sigma\|_F}{\delta}.$$

We also have

$$\sin\Theta(\hat{v}_j, v_j) \leq \frac{\|\hat{\Sigma} - \Sigma\|_{op}}{\min(|\hat{\lambda}_{j-1} - \lambda_j|, |\hat{\lambda}_{j+1} - \lambda_j|)}.$$

MATH 586 Fall 2023
Statistics for Networks

Lecture 7: Network Science: global structure

Lecturer: Robert Lunde Scribe: Vinh Pham

7.13 Sparsity

Suppose we flip a coin with probability p for each possible interaction $\{i, j\}$ or (i, j) in graph and assign an edge if heads. Then

$$\begin{aligned} E[\#\ heads] &= E[\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}] \\ &= p\binom{n}{2} \end{aligned}$$

In most of the time, real-world graphs are actually much more sparser, or we can say

$$\frac{\#\ edges}{n^2} \to 0$$

Definition 7.13.1. Sparse graphA graph is called sparse if $\frac{\#\ edges}{n^2} \to 0$ and the average degree grows with n

Definition 7.13.2. Very sparse graphThe number of edges divided by n^2 goes to 0 and the average degrees converges to a constant as n grows

Most undirected graphs have one giant component with more than 90 % of vertices, but there are graphs that have smaller components as well

7.14 Degree distribution

Most of real world networks have heavy tailed degree distributions.

Before we discuss heavy-tailed distributions, what does it mean for a distribution to have a light tail?

Definition 7.14.1. Sub-exponential distributionsA random variable X is sub $-$ Exp(K_1) if:

$$P(|X - E[X]| \geq t) \leq 2\exp(\frac{-t}{K_1}) \; \forall t \geq 0$$

Recall the union bound:

$$P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$$

and also:

$$\max_{1 \leq i \leq n} x_i > t \iff (x_1 > t) \cup (x_2 > t) \cup ... \cup (x_n > t)$$

Suppose that $X_1, \ldots, X_n \sim$ sub-exp(K_1) (which might depends on t), then we have:

$$\begin{aligned} P(\max_{1 \leq i \leq n} |X_i - E[X]| > t) &= P(\bigcup_{i=1}^{n} |X_i - E[X]| > t) \\ &\leq \sum_{i=1}^{n} P(|X_i - E[X_i]| > t) \end{aligned}$$

Now choose $t = c\log(n)$ for appropriate c . Then for n large enough,

$$\max_{1 \leq i \leq n} X_i \leq c\log(n) \;\; \text{(w.h.p)}$$

Now suppose

$$\begin{aligned} \log(f(x)) &= -a\log(x) + b \\ &= \log(x^{-a}) + b \\ &\implies f(x) = Cx^{-a} \end{aligned}$$

This implies that for some α:

$$P(X > t) = Ct^{-\alpha}$$

if $P(X > t) \sim t^{-\alpha}$ then X has Pareto tail (or "scale-free"). Recall that the Pareto distribution has $P(X > t) = (\frac{X_m}{t})^{\alpha} \;\; \forall t \geq X_m$.

For a network, α is usually in the range between 2 and 3 and for $\alpha < 2$ the second moment does not exist.

7.15 Transitivity

A common question arise is given there is an edge/connection between (i, k) and (k, j). What is the probability that there is an edge/connection from i to j

Definition 7.15.1. Transitivity coefficient

$$\begin{aligned} C &= \frac{\#\text{closed paths of length 2}}{\#\text{paths of length 2}} \\ &= \frac{6 \times \#\text{ triangles}}{\#\text{ paths of length 2}} \end{aligned}$$

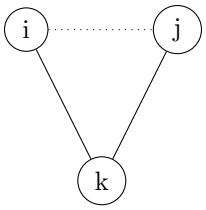


Figure 7.12: A triangle

MATH 586 Fall 2023
Statistics for Networks

Lecture 8: Lecture Title

Lecturer: Robert Lunde Scribe: Giacomo Vedovati

8.16 Global Structure of Real-World Networks

In the realm of real-world networks, several key characteristics stand out:

- **Sparsity**
- **Giant Connected Component**
- **Heavy-Tailed Degree Distributions** (commonly known as Scale-Free Networks)
- **Small World Phenomenon**
  - High transitivity + small average geodesic distances (typically around $\log n$, where n represents the number of vertices).
- **Community Structure**: These are essentially dense subnetworks within larger networks.

Understanding these features is essential, and we place particular emphasis on the significance of communities for the following reasons:

- Communities often represent functional subunits within a system.
- Social networks frequently contain distinct sub-communities.
- Analyzing network patterns becomes more insightful when considering community structure. For instance, biological networks reveal specific functions within dense subnetworks.

These elements collectively form some of the most critical aspects of network analysis.

8.17 Local Structure of Networks

One of the fundamental questions in network analysis revolves around identifying the importance of individual nodes. To elucidate this concept, consider the example of a Twitter network, where some individuals may hold more prominence, or in biological networks, certain segments may be more crucial.

Node importance can be quantified through various measures, including:

- **Degree**: For directed networks, both in-degree and out-degree are relevant.
- **Eigenvector Centrality**: Calculated as $\lambda x_i = \sum_{j \neq i} A_{ij} x_j$, which yields a higher score for nodes connected to more influential peers. For undirected networks, it simplifies to $\lambda x = A^\top x$, a result derived from the Perron-Frobenius Theorem.

Definition 8.17.1. An $n \times n$ binary matrix A is considered "irreducible" if its associated graph is strongly connected.

Theorem 8.17.2 (Perron-Frobenius)**.** *Suppose that A is an $n \times n$, irreducible matrix with spectral radius of $\rho(A) > 0$. We have the following:*

- *$\rho(A)$ is a simple eigenvalue and is positive.*
- *A possesses both positive left and right eigenvectors.*

This theorem implies that under certain conditions, the spectral radius can align with the maximum eigenvalue, which is inherently simple.

There exists a variation of the Perron-Frobenius theorem for non-negative matrices, although the corresponding eigenvectors may not strictly require positivity.

8.17.1 Pagerank Centrality

Pagerank centrality provides a scoring mechanism for nodes based on the influence they inherit from connections to highly influential nodes. It can be expressed as:

$$c_i = \alpha \sum_{j=1}^{n} \frac{A_{ji}}{d_j^{\text{out}}} c_j + 1$$

Here, α serves as a normalization factor to account for nodes' connectedness to influential counterparts. Additionally, an offset factor is included, analogous to the Perron-Frobenius theorem, which accommodates cases where irreducibility may not hold. This offset factor ensures strict positivity.

The formula can be represented in matrix form as:

$$c = \alpha A^\top D^{-1} c + I$$

Where $D = \text{diag}(d_1^{\text{out}}, \ldots, d_n^{\text{out}})$.

$$c = (I - \alpha A^\top D^{-1})^{-1}$$

8.17.2 Assortativity Coefficient

The assortativity coefficient measures the similarity between neighboring nodes. It is calculated using the formula:

$$r_x = \frac{\sum_{(i,j) \in E)} (X_i - \bar{x})(X_j - \bar{x})}{\sqrt{\sum_{(i,j) \in E)} (X_i - \bar{x})^2 \sum_{(i,j) \in E)} (X_j - \bar{x})^2}}$$

Here, $\bar{X}$ represents the average of attribute values associated with nodes.

The coefficient provides a correlation measure that can be computed for various networks. The sum of individual assortativity coefficients (r_i) across nodes gives an overall network assortativity (r).

In essence, the assortativity coefficient helps gauge how closely related neighboring nodes are within a network.

MATH 586 Fall 2023
Statistics for Networks

Lecture 9: Network Sampling I

Lecturer: Robert Lunde Scribe: Anthony Hong

Before staring the new topic of sampling, we will first finish some last parts about local structures of network.

9.18 Last Part of Local Structures of Networks

9.18.1 Measures of Node Importance

Many have been proposed, here are some commonly used.

-Eigenvector Centrality

-Page Rank Centrality

-Closeness Centrality

$$C_i = \frac{1}{n}\sum_{i=1}^{n} d(i,j), \text{where } d(,) \text{ is geodistic distrance}$$

-Local Transitivity

$$C_i = \frac{\#\text{ neighbours of i that sharing an edge}}{\#\text{ pairs of neighbours}}$$

$$C = \sum_{i=1}^{n} C_i$$

Remark: this measure is different from transitivity.

9.18.2 Measure of Similarity

-Assortativity Coefficient (e.g. assortative by degree)

-Common Neighbours

$$S(i,j) = \sum_{k=1}^{n} A_{ik}A_{jk}$$

$$S^{'}(i,j) = \frac{\sum_{k=1}^{n} A_{ik}A_{jk}}{\sqrt{d_i d_j}}$$

9.19 Introduction

When we study the network structure of some objects, it is often the case that modelling the whole collection of objects in a single group is impossible due to unavailability of the data and lack of ability to handle this large graph computationally. We then consider sampling this collection to can infer the structure by the samples. Let $G = (V, E)$ be a **population graph** that includes the all from which we sample. We take a sample of nodes $E^* \subseteq E$ and edges $V^* \subseteq V$ to form a **sample graph** $G^* = (V^*, E^*)$. Suppose we have some graph parameter $\eta(G)$. Can we learn this parameter from our samples by building an estimator $\hat{\eta}$ from G^*? What are some common network sampling schemes?

Our experience in classical statistical point estimation tells us we may just build $\hat{\eta} = \eta(G^*)$, but it turns out that this does not quite work in the graph setting. [1] example 5.1 illustrates the subtlety.

9.20 Network Sampling Schemes

We give some examples of commonly used network sampling schemes.

1 **Induced subgraph sampling/node sampling**: Take a random sample of n vertices from the population graph G and observe the subgroup induced by these vertices.

2 **Incident subgraph sampling/edge sampling**: Take a random sample of m edges from the population graph G and observe the subgroup with all of these edges and vertices incident to them. This is a dual version of node sampling.

3 **Ego Sampling**: Observe all the nodes that are neighbors of a fixed vertex and often complete them by the induced subgraph.

4 **Snowball sampling**: Pick a subset of the vertex set $V_0 \subseteq V$ as the **seed nodes**. In the first wave, we include neighbors of seed nodes that aren't themselves seeds, resulting in $V_1 \supseteq V_0$ Inductively, in the k-th wave, we include neighbors of V_{k-1} that have not appeared in the sample. This sampling process typically continues until no new nodes recruited or some stopping criteria met. Note that the number of recruits can also be a fixed number. This sampling is closely related to respondent-driven data (RDS).

9.21 Sampling Bias

Now we come back to our original inference question. How can we account for the sampling bias associated with these different schemes? Suppose we have a population $\mathcal{U} = \{1, \cdots, N\}$, each unit $i \in \mathcal{U}$ associated with a scalar value of interested y_i. The **inclusion probability** of the unit i is the probability of inclusion of the unit i in any sample with respect to the sampling design D and will be denoted by π_i. We then look at a simple estimation case: the estimation for the population mean

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i$$

When sampling with unequal probabilities (that is, there are higher chances of picking some over others), the sample mean can be biased. The **Horvitz-Thompson estimator** solves this problem via weighted mean. Let S be the indices of a sample of vertices of size n. Let π_i be the probability of including node i in

the sample. Assuming $\pi_i > 0$ for each unit i, the estimator is given by

$$\hat{\mu} = \frac{1}{N}\sum_{i\in S}\frac{y_i}{\pi_i}$$

Theorem 9.21.1. *Horvitz-Thompson estimator is an unbiased estimator, assuming $\pi_i > 0$.*

Proof. Letting z_i be the Bernoulli variable indicating whether node i belongs to the S, We calculate the expectation

$$\begin{aligned}\mathbb{E}(\hat{\mu}) &= \mathbb{E}\left(\frac{1}{N}\sum_{i\in S}\frac{y_i}{\pi_i}\right) = \mathbb{E}\left(\frac{1}{N}\sum_{i\in\mathcal{U}}\frac{y_i z_i}{\pi_i}\right)\\ &= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{y_i z_i}{\pi_i}\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{E}(z_i)\\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\pi_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu\end{aligned}$$

where we used the fact $\mathbb{E}(z_i) = \mathbb{P}(z_i = 1) = \pi_i$ □

MATH 586 Fall 2023
Statistics for Networks

Lecture 10: Network Sampling II and Markov Chain I

Lecturer: Robert Lunde Scribe: Anthony Hong

10.22 Horvitz-Thompson Estimator (cont'd)

Let $G = (V, E)$ be the population graph and $G^* = (V^*, E^*)$ be the sample graph and $\eta(G)$ be the graph parameter of interest. Let the population be $\mathcal{U} = \{1, \cdots, N\}$, each unit $i \in \mathcal{U}$ associated with a scalar value of interested y_i, and let μ be the population mean. We continue our study of network sampling. Recall the **Horvitz-Thompson estimator** is given by

$$\hat{\mu} = \frac{1}{N}\sum_{i\in S}\frac{y_i}{\pi_i}$$

where S is the indices of a sample of vertices of size n, π_i be the probability of including node i in the sample. Letting z_i be the Bernoulli variable indicating whether node i belongs to the S, we notice that $\mathbb{E}(z_i z_j) = \pi_{ij}$ (the probability of including both) and $\mathbb{E}(z_i) = \pi_i$ and $\mathbb{E}(z_j) = \pi_j$. Then the variance is given by

$$\begin{aligned}\mathrm{Var}(\hat{\mu}) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i\in\mathcal{U}}\frac{y_i z_i}{\pi_i}\right)\\ &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}}\frac{y_i y_j}{\pi_i\pi_j}\mathrm{Cov}(z_i, z_j)\\ &= \frac{1}{N^2}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}} y_i y_j\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)\end{aligned}$$

and we have the estimator

$$\hat{\sigma}^2_{\hat{\mu}} = \frac{1}{N^2} \sum_{i \in S} \sum_{j \in S} y_i y_j \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}} \right)$$

so that $\mathbb{E}(\hat{\sigma}) = \text{Var}$

We can also compute Horvitz-estimate for parameter of the form

$$\mathcal{T} = \sum_{(i,j) \in \mathcal{U}^2} y_{ij}$$

where y_{ij} is the value of interest for each edge (i, j). The estimator is given by

$$\hat{\mathcal{T}} = \sum_{(i,j) \in S^2} \frac{y_{ij}}{\pi_{ij}}$$

We end this section with estimation of the number of vertices N on the graph, or the total population. One common method for estimating N is called "capture-recapture". Consider a sampling without replacement with two stages:

(1) Mark all vertices in first stage to get S_1;

(2) In the second stage, count how many vertices re-appear.

Calculate the estimator

$$\hat{N} = \frac{n_2}{m} n_1$$

where $|S_1| = n_1$, $|S_2| = n_2$, and m is the number of vertices re-appear in the second stage. To see why this estimator works, we see this sampling gives a hypergeometric distribution below

$$L(m; N) = \frac{\binom{n_1}{m} \binom{N-n_1}{n_2-m}}{\binom{N}{n_1+n_2}}$$

It is an exercise to see that the MLE of N is then $\hat{N}$.

10.23 Stochastic Process: Markov chain on discrete space

A **stochastic process** is a collection of E-valued random variables $\{X_t : \Omega \to E\}_{t \in T}$ with **state space** $(E, \mathcal{E})$ and **parameter set** T. For each $\omega \in \Omega$, let $X(\omega)$ denote the function $T \to E; t \mapsto X_t(\omega)$; then $X(\omega)$ is an element of E^T, the collection of all functions from T to E. We may regard the stochastic process $(X_t)_{t \in T}$ as a random variable X that takes value in the product space $(E^T, \mathcal{E}^T)$, since the map $X : \Omega \to E^T; \omega \mapsto X(\omega)$ is measurable relative to $\mathcal{D}$ and $\mathcal{E}^T$. When $T = \mathbb{N} = \{0, 1, \cdots\}$ and $(E, \mathcal{E}) = (I, \mathcal{I})$, we come to the definition of the first major class of stochastic process in this note, adapted from [2]:

Definition 10.23.1 (Discrete-Time Markov Chain)**.** We say $(X_n)_{n \in \mathbb{N}}$ is a **discrete-time Markov chain** with **initial distribution** λ and **transition matrix** $P = (p_{ij})_{i,j \in \mathbb{N}}$, or **Markov**$(\lambda, P)$ for short, if

(i) X_0 has distribution λ, i.e., $\forall i_0 \in I : \mathbb{P}(X_0 = i_0) = \lambda_{i_0}$

(ii) for $n \geq 0$, conditional on $X_n = i$, X_{n+1} has distribution $(p_{ij})_{j \in I}$, i.e., $\mathbb{P}(X_{n+1} = j | X_n = i) = p_{ij}$;

(iii) for $n \geq 0$, $(X_{n+1} = j | X_n = i)$ is independent of $X_0, \cdots, X_{n-1}$. ♦

We leave it as an exercise by induction to show that (iii) is equivalent of saying

$$\forall i_0, \cdots, i_{n+1} \in I : \mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, \cdots, X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n)$$

Continuous-time Markov chain $(X_t)_{t\geq 0}$ is a stochastic process with t belonging to an uncountable parameter set $T = [0, \infty)$ and random variables $X_t : \Omega \to I$, where I is still a countable set. Formal treatment of this requires us to understand Q-matrices (see [2] chapter 2 and 3). We continue our study of the discrete version. Here are two properties of it.

Proposition 10.23.2.

(i) $(X_n)_{0\leq n\leq N}$ *is Markov*(λ, P) *if and only if*

$$\forall i_0, \cdots, i_N \in I : \mathbb{P}(X_0 = i_0, \cdots, X_1 = i_1, \cdots, X_N = i_N) = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{N-1} i_N}$$

(ii) ***(Markov property)****: Let* $(X_n)_{n\geq 0}$ *be Markov*(λ, P)*. Then, conditional on* $X_m = i$*,* $(X_{n+m})_{n\geq 0}$ *is Markov*(δ_i, P) *and is independent of the random variables* $X_0, \cdots, X_m$*, where* $\delta_i = (\delta_{ij} : j \in I)$ *is the* ***unit mass*** *at* i *and*

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

We regard distributions and measures λ as row vectors whose components are indexed by I, just as P is a matrix whose entries are indexed by $I \times I$. We extend the matrix multiplication and multiplication of matrix by row vector to the general sense in the obvious way, defining a new measure λP and a new matrix P^2 by

$$(\lambda P)_j = \sum_{i\in I} \lambda_i p_i j, \quad (P^2)_{ij} = \sum_{k\in I} p_{ik} p_{kj}.$$

We define P^m similarly for any n. We set $P^0 = I$ where $(I)_{ij} = \delta_{ij}$. We write $p_{ij}^{(n)} = (P^n)_{ij}$ for the (i, j) entry in P^n. In the case where $\lambda_i > 0$ we shall write $\mathbb{P}_i(A)$ for the conditional probability $\mathbb{P}(A|X_0 = i)$. By the Markov property at time $m = 0$, under $\mathbb{P}_i$, $(X_n)_{n\geq 0}$ is Markov(δ_i, P). So the behaviour of $(X_n)_{n\geq 0}$ under $\mathbb{P}_i$ does not depend on λ.

Theorem 10.23.3 ([2] Theorem 1.1.3)**.** *Let* $(X_n)_{n\geq 0}$ *be Markov*(λ, P)*. Then, for all* $n, m \geq 0$*,*

(i) $\mathbb{P}(X_n = j) = (\lambda P^n)_j$*;*

(ii) $\mathbb{P}_i(X_n = j) = \mathbb{P}(X_{n+m} = j | X_m = i) = p_{ij}^{(n)}$

Proof. (i) By Proposition 2.2 (i)

$$\begin{aligned} \mathbb{P}(X_n = j) &= \sum_{i_0\in I} \cdots \sum_{i_{n-1}\in I} \mathbb{P}(X_0 = i_0, \cdots, X_{n-1} = i_{n-1}, X_n = j) \\ &= \sum_{i_0\in I} \cdots \sum_{i_{n-1}\in I} \lambda_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} j} = (\lambda P^n)_j \end{aligned}$$

(ii) By the Markov property, conditional on $X_m = i$, $(X_{m+n})_{n\geq 0}$ is Markov(δ_i, P); then take $\lambda = \delta_i$ in (i). □

In light of this theorem we call $p_{ij}^{(n)}$ the **n-step transition probability from** i **to** j. The following examples illustrates how to calculate it.

Example 10.23.4 ([2] Example 1.1.4)**.** Suppose that whether it rains tomorrow depends on previous weather conditions only through whether it is raining today. Suppose further that if it is raining today, then it won't rain tomorrow with probability α, and if it is not raining today, then it will rain tomorrow with probability β. If we say that the system is in state 0 when it rains and state 1 when it does not, then the preceding system is a two-state Markov chain having transition probability matrix

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

and is represented by Figure 3.

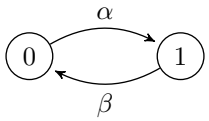


Figure 10.13: A two-state Markov chain

We exploit the relation $P^{n+1} = P^n P$ to write

$$p_{00}^{(n+1)} = (P^{n+1})_{00} = (P^n P)_{00} = \sum_{k=0,1} p_{0k}^{(n)} p_{k0} = p_{00}^{(n)} p_{00} + p_{01}^{(n)} p_{10} = p_{00}^{(n)}(1-\alpha) + p_{01}^{(n)}\beta$$

We also know that $p_{00}^{n} + p_{01}^{(n)} = \mathbb{P}_0(X_n = 0 \text{ or } 2) = 1$, so by eliminating $p_{01}^{(n)}$ we get a recurrence relation for $p_{00}^{(n)}$:

$$p_{00}^{(n+1)} = p_{00}^{(n)}(1-\alpha-\beta) + \beta, \quad p_{00}^{(0)} = 1.$$

This has a unique solution:

$$p_{00}^{(n)} = \begin{cases} \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n, & \text{for } \alpha+\beta > 0, \\ 1, & \text{for } \alpha+\beta = 0. \end{cases} \qquad \Diamond$$

As illustrated in the scenario above, we percieve $X = (X_n)_{n\geq 0}$ as a kind of "random walker" on a graph with $|I|$ vertices ($|I| = 2$ in this case). The example only specifies the P matrix, but a complete discrete-time Markov chain also includes an intial distirbution by which the random variable X_0 assigns the position i_0 where the walker is born. X_1 assigns i_1 where the walker goes from i_0, $\cdots$, X_{n+1} assigns i_{n+1} where the walker goes from i_n. p_{ij} is the probability that X will go to position j given X is currently at position i, but notice that X's choice is regardless of the time or step when X is at. Namely,

$$\forall i_0, \cdots, i_{n+1} \in I : \mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, \cdots, X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n) = p_{ij}$$

MATH 586 Fall 2023
Statistics for Networks

Lecture 11:Random Walks Continued, Part II

Lecturer: Robert Lunde Scribe: Samuel Naranjo Rincon

11.24 Introduction

A random walk on a graph is a time-homogeneous discrete time markov chain with n states, where the transition probability is given by:

$$P(X_{n+1} = j | X_n = i) = \begin{cases} \frac{1}{d_i} & A_{ij} = 1 \\ 0 & o.w. \end{cases}$$

So, the probability of a ste, or edge ij, is given by the conditional probability $P(X_{n+1} = j | x_n = i)$. And this is influenced by the degree of node i.

11.25 Markov Properties:

We have that Markov chains hold this probability property:

$$P(X_{n+1} = \lambda_{n+1} | (X_n = \lambda_n, ..., (X_0 = \lambda_0) = P((X_{n+1} = \lambda_{n+1} | (X_n = \lambda_n)$$

This Markov property implies:
$P(X_n = \lambda_n, .., X_o = \lambda_0)$
$= P(X_n = \lambda_n | X_{n-1} = \lambda_{n-1}, ..., X_0 = \lambda_0) * P(X_{n-1} = \lambda_{n-1}, ..., X_0 = \lambda_0)$
$= P(X_n = \lambda_n | X_{n-1} = \lambda_{n-1}) * P(X_{n-1} = \lambda_{n-1} | X_{n-2} = \lambda_{n-2}) * ... * P(X_0 = \lambda_0)$

Proposition:
$P_{ij}^m = P(X_m = j | X_0 = i)$

Proof for m=2:

$$P(X_2 = j | X_0 = i) = \sum_{k=1}^{w} P(X_2 = j | X_1 = k, X_0 = i) P(X_1 = k | X_0 = i)$$

$$= \sum_{k=1}^{n} P(X_2 = j | X_1 = k) P(X_1 = k | X_0 = i) = \sum_{k=1}^{n} P_{ik} P_{kj} = P_{ij}^2$$

Let $\pi_0 = (P(X_0 = 1), ..., P(X_o = n))$

Proposition:
$P(X_m = i) = (\pi_0^{'} P^m)i$
$P(X_1 = i) = \sum_{k=1}^{w} P(X_1 = i|X_0 = k)P(X_0 = k)$
$= \sum_{k=1}^{w} \pi_0(k)P_{ki}$
$= (\pi_0^{'} P^m)i$

- $(X_t)_{t\geq 1}$ is stationary if from any n, $t_1, ..., t_n \geq 1$ and for any t: $(X_{k1}, ..., X_{kn}) \stackrel{d}{=} (X_{t1+\tau}, ..., X_{tn+\tau})$
- A Markov chain is stationary if π_0 satisfies:

$$\pi_0^{'} p = \pi_0^{'}$$
$$\pi_0^{'} p^2 = (\pi_0^{'} p)p$$
$$= \pi_0^{'} p$$
$$= \pi_0^{'}$$

- A stochastic process X_t where $t \geq 1$ is time reversible if for any n, t_n and τ: $(X_t, ..., x_{tn}) \stackrel{d}{=} (X_{\tau - t1}, ..., X_{\tau - tn})$

- A Markov chain is reversible if the detailed balance condition, given by: $\lambda(i)P_{ij} = \lambda(j)P_{ji} \forall$ holds.

Proposition:
If detailed balance condition holds, (X_t) where $t \geq 1$ is stationary.
Proof:

$$\stackrel{n}{_{k=1}} \lambda(k)P_{kj} = \sum_{k=1}^{n} \lambda(j)P_{jk}$$
$$\implies \sum_{k=1}^{n} \lambda(k)P_{kj} = \lambda(j) \sum_{k=1}^{n} P_{jk}$$
$$\implies \sum_{k=1}^{n} P(X_1 = j|X_0 = k)P(X_0 = k) = \lambda(j)$$
$$\implies P(X_1 = j) = \lambda(j)$$

Proposition:
$\lambda(i) = \frac{d_i}{2|E_i|}$ is a stationary distribution for a random walk on a graph.
Proof:
Plug into a detailed balance:

$$\frac{d_i}{2|E_i|}\frac{1}{d_i} = \frac{d_i}{2|E_j|}\frac{1}{d_j}$$

if $A_{ij} = 1$

Question: When does $\lim_{k\to\infty} P^t = \pi$ where $\pi = \begin{bmatrix}\lambda\\ \lambda\\ \vdots\\ \lambda\end{bmatrix}$

For random walks on graphs, this holds when graph is connected and not bipartite.
Observe that $P = D^{-1}A$
It also turns out that P is similar to $A = D^{-1/2}AD^{-1/2}$
This is because we know that two matrices A and B are similar if there exists matrix C such that: $A = CBC^{-1}$
We start with that and show the following:

$$P = D^{-1/2}AD^{-1/2}$$

$$P^t = D^{-1/2}A^tD^{-1/2}$$

$$= D^{-1/2}\sum_{k=1}^{n}\lambda^t A v_k v_k^{'} D^{-1/2}$$

MATH 586 Fall 2023
Statistics for Networks

Lecture 12:Fundamental Theorem of Markov Chains

Lecturer: Robert Lunde Scribe: Sidney Nwakanma

12.26 Introduction

Fundamental Theorem of Markov Chain is a general result that establishes conditions under which

$$\lim_{t\to\infty} \boldsymbol{P^t} = \boldsymbol{\Pi},$$

where

$$\boldsymbol{\Pi} = \begin{bmatrix}\pi\\ \pi\\ .\\ .\\ .\\ \pi\end{bmatrix}$$

In otherwords, $\forall i, j$, $\lim_{t\to\infty} \boldsymbol{P}_{ij}^t = \boldsymbol{\pi(j)} = \boldsymbol{P(x_0 = j)}$.

For a Random Walk on a graph, $\lim_{t\to\infty} \boldsymbol{P}^t = \boldsymbol{\Pi}$, if G is connected and not bipartite.

$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A}$ (Random Walk Laplacian)

$\boldsymbol{P} = \boldsymbol{D}^{-\frac{1}{2}}\mathcal{A}\boldsymbol{D}^{-\frac{1}{2}}$ where $\mathcal{A} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ (normalized adjacency matrix).

$\boldsymbol{P}^t = \boldsymbol{D}^{-\frac{1}{2}}\mathcal{A}\boldsymbol{D}^{-\frac{1}{2}}$

$\boldsymbol{P}^t = \sum_{k=1}^n \boldsymbol{D}^{-\frac{1}{2}}\lambda_k^t \nu_k \nu_k^{'}\boldsymbol{D}^{-\frac{1}{2}}$ and recall that $x^{'}\alpha x = \sum_{(i,j)\in E}\left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}}\right)$

$\boldsymbol{P}_{ij}^t = \sum_{k=1}^n \lambda_k^t \nu_k(i)\nu_k^{'}(j)\sqrt{\frac{d_i}{d_j}}$

$\boldsymbol{P}_{ij}^t = \mathbf{1}^t\sqrt{\frac{d_i d_j}{2|E|2|E|}}\sqrt{\frac{d_i}{d_j}} + \sum_{k=1}^n \lambda_k^t \nu_k(i)\nu_k^{'}(j)\sqrt{\frac{d_i}{d_j}}\ with\ |\lambda| < 1$

$\boldsymbol{P}_{ij}^t = \frac{d_j}{2|E|} + R\ where\ \frac{d_j}{2|E|} = \pi(j)$

12.27 Lemma:

Suppose that G is bipartite. Then $\lambda_{min} = -1$ if and only if G is bipartite.

If λ is an eigenvalue of $\boldsymbol{A}$ then so is $-\lambda$.

Proof:

If G is bipartite, then:

$$A = \begin{bmatrix} 0 & B \\ B' & 0 \end{bmatrix}$$

Suppose $\begin{bmatrix} x \\ y \end{bmatrix}$ is eigenvector of $\boldsymbol{A}$ with value λ

$$\boldsymbol{A}\begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix}$$

so

$$By = \lambda x$$
$$B^{'}x = \lambda y$$

Consider $\begin{bmatrix} x \\ y \end{bmatrix}$ where

$$\boldsymbol{A}\begin{bmatrix} x \\ -y \end{bmatrix} = \begin{bmatrix} -By \\ B^{'}x \end{bmatrix} = \begin{bmatrix} -\lambda x \\ \lambda y \end{bmatrix} = -\lambda \begin{bmatrix} x \\ -y \end{bmatrix}$$

Proposition:

Suppose G is connected. Then $\lambda_{min} = -1$ iff G is bipartite.

Proof:

$$
\begin{aligned}
|\lambda_{min}| &= |\sum_{i,j} \boldsymbol{A}_{i,j}\nu_i nu_j| \; where \; ||\boldsymbol{\nu}|| = 1 \\
&\leq \sum_{i,j} |A_{i,j}||\nu_i||\nu_j| \\
&\leq \sum_{i,j} |A_{i,j}| z_i \cdot z_j
\end{aligned}
$$

where Z is eigenvectors corresponding to λ_{max}

For equality to hold, it must be the case that $|\nu_i\nu_j| = Z_i \cdot Z_j$, $\forall (i,j) \in E$

For each edge, one coordinate must be positive, the other negative.

We can partition nodes into the following sets: $U = \{i : v_i > 0\}$ and $V = \{j : v_j < 0\}$, since no edges can be present within U an V. G connected and $\lambda_{min} = -1$, then G is Bipartite.

Let $\Delta(t) = max_{i,j}|\frac{P_{ij}}{\lambda_j} - 1|$ we can show that $\Delta(t) = \frac{\gamma^t}{\lambda_{min}}$ where $\gamma = max(\lambda_2, |\lambda_n|)$

Proof:

$$
\begin{aligned}
&\Delta(t) = max_{i,j}|\frac{P_{ij}}{\lambda_j} - 1| \\
&\leq max_{i,j} \sum_{k=2}^{n} \frac{|\lambda_k|^t |V_k(i)||V_k(j)|}{\lambda_j} \cdot \sqrt{\frac{d_i}{d_j}} \\
&\leq max_{i,j}\gamma^t \sum_{k=1}^{n} \frac{|V_k(i)||V_k(j)|}{\sqrt{\lambda_i\lambda_j}} \; where \; \lambda(i) = \frac{d_i}{2|E|} \\
&\leq max_{i,j}\gamma^t \frac{(\sum_{k=1}^{n} |V_k^2(i)|)(\sum_{k=1}^{n} |V_k^2(j)|)}{\lambda_{min}} \; where \; ||Vu|| = 1 \\
&= \frac{\gamma^t}{\lambda_{min}}
\end{aligned}
$$

MATH 586 Fall 2023
Statistics for Networks

Lecture 13: Random Walks in Graph IV

Lecturer: Robert Lunde Scribe: Aaron Luo

13.28 Introduction

A random walk on a graph is a Markov chain where:

$$
P_{ij} = P(X_{n+1} = j \mid X_n = i) = \begin{cases} \frac{1}{d_i} & \text{if } A_{ij} = 1 \\ 0 & \text{otherwise} \end{cases}
$$

The stationary distribution π is given by:

$$\pi = \left(\frac{D_1}{2|\mathcal{E}|}, \ldots, \frac{D_n}{2|\mathcal{E}|} \right)$$

If G is connected and not bipartite, then we have:

$$\lim_{t \to \infty} P_{ij}^t = \pi_j, \text{ where } \pi = \left(\frac{1}{\sqrt{D_i}}, \ldots, \frac{1}{\sqrt{D_n}} \right)^T$$

For a lazy random walk on the graph:

$$Q = \frac{1}{2} I + \frac{1}{2} P$$

Define the discrepancy as:

$$\Delta(t) = \max_{i,j} \left| P_{ij}^t - \pi_j \right|$$

For a random walk (RW) on a graph, we have:

$$\Delta(t) \leq \frac{\gamma^t}{\pi_{\min}}$$

where

$$\gamma = \max_{i,j} (D_i, D_j)$$

and

$$\pi_{\min} = \min (\pi_1, \ldots, \pi_n)$$

For two probability mass functions P and Q, the total variation distance is defined as:

$$\|P - Q\|_{\text{TV}} = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$$

For discrete distributions, it is clear that

$$\max_{A \in \mathcal{A}} P(A) - Q(A)$$

is attained by

$$A = \{x : P(x) \geq Q(x)\}$$

Using similar arguments,

$$\max_{A \in \mathcal{A}} Q(A) - P(A)$$

is attained by

$$A^c = \{x : P(x) < Q(x)\}$$

The total variation distance between two probability mass functions P and Q is:

$$\|P - Q\|_{\text{TV}} = \frac{1}{2} \sum_{x \in S} |P(x) - Q(x)|$$

This can be rewritten as:

$$\|P - Q\|_{\text{TV}} = \frac{1}{2}\left(\sum_{x \in A} |P(x) - Q(x)| + \sum_{x \in A^c} |P(x) - Q(x)|\right)$$

where A and A^c are defined as above.

We are often interested in:

$$\sup_{x \in \mathcal{X}} \|P^t(\cdot \mid X_0 = i) - \pi\|_{\text{TV}}$$

The total variation distance is given by:

$$\|P^t(\cdot \mid X_0 = i) - \pi\|_{TV} = \frac{1}{2}\sum_{j=1}^{n} \left|\frac{P_{ij}^t}{\pi_j} - 1\right| \pi_j$$

This can be further bounded by:

$$\leq \frac{1}{2}\sum_{j=1}^{n} \max_{i,j} \left|\frac{P_{ij}^t}{\pi_j} - 1\right| \pi_j$$

And simplified to:

$$= \Delta(t)\frac{1}{2}\sum_{j=1}^{n} \frac{\pi_j}{\pi_j} = \Delta(t)$$

The total variation distance can be expressed as:

$$\|P^t(\cdot \mid X_0 = i) - \pi\|_{TV} = \frac{1}{2}\sum_{j=1}^{n} \left|\frac{P_{ij}^t}{\pi_j} - 1\right| \pi_j$$

This can be further bounded by:

$$\leq \frac{1}{2}\sum_{j=1}^{n} \max_{i,j} \left|\frac{P_{ij}^t}{\pi_j} - 1\right| \pi_j$$

And simplified to:

$$= \Delta(t)\frac{1}{2}\sum_{j=1}^{n} \frac{\pi_j}{\pi_j} = \Delta(t)$$

For the supremum over i we have:

$$\sup_{i} \|P^t(\cdot \mid X_0 = i) - \pi\|_{TV} \leq \frac{1}{\sqrt{\pi_{\min}}}\gamma^t$$

Conductance

The conductance of the cut (S, S^c) is defined as:

$$\phi(S, S^c) = \frac{|\partial(S)|}{\min(d(S), d(\overline{S}))}$$

where

$$\partial(S) = \{(i, j) \in E \mid i \in S, j \in S^c\}$$

and

$$d(S) = \sum_{i \in S} d_i$$

Recall

$$\pi(S) = \sum_{i \in S} \frac{d_i}{2|E|}$$

Thus, we have

$$\phi(S, S^c) = \frac{|\partial(S)|}{2|E|\pi(S)}$$

Observe

$$\frac{|\partial(S)|}{2|E|} = \sum_{i \in S} \sum_{j \in S^c} \frac{\pi(i) P_{ij}}{\pi(S, S^c)}$$

since

$$\pi(S, S^c) = \sum_{i \in S} \sum_{j \in S^c} \pi(i) P_{ij}$$

$$\sum_{i \in S} \sum_{j \in S^c} \frac{d_i}{2|E|} P_{ij} = \sum_{i \in S} \sum_{j \in S^c} \frac{d_i}{2|E|} \frac{1}{d_i} \mathbf{1}(A_{ij} = 1)$$

Which simplifies to:

$$= \frac{|\partial(S)|}{2|E|}$$

which is a tighter bound.

MATH 586 Fall 2023
Statistics for Networks

Lecture 14: Random Walks in Graph IV

Lecturer: Robert Lunde Scribe: Adrian Cao

14.29 Random Walks on graph Review

14.29.1 Matrix representations

$$P_{ij} = P(X_{n+1} = j | X_n = i) = \begin{cases} \frac{1}{d_i} & \text{if } A_{ij} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

With this transition matrix, a stationary distribution is given by

$$\pi = (\frac{d_1}{2|E|}, \ldots, \frac{d_n}{2|E|})$$

14.29.2 Total Variaton norm

$$\|P - Q\|_{TV} = \sup_{A \in \mathcal{A}} |P(A) - Q(A)| = \frac{1}{2} \sum_{X \in S} |p(x) - q(x)|$$

The second equality holds if p,q are discrete

For random walks on the graph:

$$\sup \|P^t(\cdot | x_0 = i) - \pi\|_{TV} \leq \frac{\gamma}{\pi_{min}} \textbf{ where } \gamma = \max(\lambda_2, |\lambda_n|), \pi_{min} = \frac{\min_i d_i}{2|E|}$$

14.29.3 Conductance of a cut

$$\phi(S, \bar{S}) = \frac{|\partial(S)|}{\min(d(S), d(\bar{S})}$$

where $\partial(S) = \{(u, v) \in E : u \in S, v \in \bar{S}\}$, $d(S) = \sum_{i \in S} d_i$

Conductance of graph

$$\phi_G = \min_S \phi(S, \bar{S}) = \min_{|S| \leq \frac{|v|}{2}} \frac{|\partial(S)|}{d(S)} = \min_{|S| \leq \frac{|V|}{2}} \frac{\pi(S, \bar{S})}{\pi(S)}$$

Another notion of conductance

$$\Phi(S, \bar{S}) = \frac{\pi(S, \bar{S})}{\pi(S)\pi(\bar{S})}$$

$$\Phi_G = \min_S \Phi(S, \bar{S})$$

Theorem 14.29.1 (Cheeger Inequality)**.**

$$\frac{\phi_G^2}{2} \leq 1 - \lambda_2(A) \leq 2\phi_G, \frac{\Phi_G^2}{8} \leq 1 - \lambda_2(A) \leq \Phi_G$$

Suppose that G is a directed graph, we could write $P = D^{-1}A$, but connectness is not guaranteed. Then, we could transform $P = (1 - \alpha)D^{-1}A + \alpha B$ to make it connected. Here, we use $d_i^o ut = \sum_{j=1}^n A_{ij}$ and $B = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$.

14.30 Metropolis-Hasting Algorithm

Recall for the previous random walks in graph, we have $\pi = (\frac{D_1}{2|E|}, \cdots, \frac{D_n}{2|E|})$.

Suppose we want to modify a Markov Chain to have a different stationary measure. Let $P(\cdot|\cdot)$ be the transition kernel associated with the original Markov chain. Consider the following algorithm.

Algorithm 1 Metropolis-Hasting for Random Walks in Graph

1: Generate $Y_{n+1} \sim P(\cdot|X_n = x_n)$

2: $X_{n+1} = \begin{cases} y_{n+1} & \text{with probability } \rho(X_n, X_{n+1}) \\ x_n & \text{with probability } 1 - \rho(X_n, X_{n+1}) \end{cases}$

3: $\rho(x, y) = \min(\frac{\phi(y)P(x|y)}{\phi(x)P(y|x)}, 1)$

Show that π is stationary distribution using detailed balance $\pi_i Q_{ij} = \pi_j Q_{ji} \forall i, j$.

$$\pi_i P_{ij} \frac{\pi_j P_{ji}}{\pi_i P_{ij}} = \pi_j P_{ji}$$

Example, suppose we want to have the stationary distribution of interest is $\pi = (\frac{1}{n}, \cdots, \frac{1}{n})$, then we just want to have $\rho(x, y) = \min(\frac{d_x}{d_y}, 1)$, and $Q_{ij} = \begin{cases} \frac{1}{d_i} & \text{if } d_i \leq d_j \\ \frac{1}{d_j} & \text{if } d_j > d_i \end{cases}$

MATH 586 Fall 2023
Statistics for Networks

Lecture 15: Random Graph

Lecturer: Robert Lunde Scribe: Sayan Das

15.31 Random Graph models

- *Starting point*: some notion of uniform distribution on graphs.
- $G_{n,m}$ is a model on nodes where all graphs with m edges equally likely,

$$P(G = g) = \frac{1}{\binom{\binom{n}{2}}{m}}, \text{ for all } g \text{ with } m \text{ edges.}$$

- $G_{n,p}$ *model*: assign to each graph g with n nodes and m edges.

$$P(G = g) = p^m (1 - p)^{\binom{n}{2} - m}$$

  Equivalently, $A_{ij} = A_{ji} \sim Ber(p)$. For $G_{n,p}$, it is immediate that $D_i \sim Bin(n - 1, p)$, $D_i = \sum_{j \neq i} A_{ij}$.

- It is common study $G_{n,p}$ with $p \to 0$.

- *Key questions*: what are thresholds for certain graph properties to appear?
  For example, it can be shown that

$$\lim_{n\to\infty} P(G_n \text{ is connected}) = \begin{cases} 1, & c > 1 \\ 0, & c < 1 \end{cases}, \; p_n = \frac{c\log(n)}{n}.$$

- It can be shown

$$P(G \text{ contains isolated node}) \to 1, \quad \text{if } p_n < \frac{\log n}{n}$$

- It will often be the case that $G_{n,p}$ and $G_{n,m}$ have analogous properties if one considers $p = \frac{m}{\binom{n}{2}}$.
  Idea: Some threshold where giant components emerge. Let $p_n = \frac{\lambda}{n}$, if $\lambda > 1$: giant component; $\lambda < 1$: multiple smaller components.

Theorem: *Suppose $\lambda > 1$, then $\exists$ a constant B such that with probability tending to 1, there is only one component with more than $B\log n$ vertices. Moreover, the size of the largest component is $\Theta(n)$.*

Theorem: *Suppose $p = \lambda/n,\ \lambda < 1$. For all sufficiently large a,*

$$P\Big(\max_{1\le i\le n} |S_i| > a\log n\Big) \to 0$$

where $|S_i|$ is the size of component that contains node i.

MATH 586	Fall 2023
Statistics for Networks	
Lecture 16:	
Lecturer: Robert Lunde	Scribe: Wei Li

16.32 Exponential Random Graph Models

Definition: *Suppose $(X_1,\dots,X_n) \sim P_\theta$, where $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. A statistic $T((X_1,\dots,X_n))$ is sufficient for θ if the distribution of $(X_1,\dots,X_n)|T$ doesn't depend on θ.*

Theorem: *A statistic $T(X_1,\dots,X_n)$ is sufficient for θ if and only if the joint pdf/pmf permits a factorization of the form,*

$$f(x_1,\dots,x_n) = h(x_1,\dots,x_n)\; g(T,\theta).$$

Exponential family: *A class of distribution $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is an exponential family if pdf/pmf permits factorization of the form,*

$$f(x;\theta) = h(x)\ \exp(\eta(\theta)^T T(x) - A(\theta)) = h(x_1,\dots,x_n)\; g(T,\theta) \textit{(Factorization form).}$$

Example 1: $Poi(\lambda)$

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{1}{x!}\exp x\ \log(\lambda) - \lambda.$$

Example 2: $Bin(n,p)$

$$P_X(x) = \binom{n}{x}p^x(1-p)^{n-x} = \binom{n}{x}\Big(\frac{p}{1-p}\Big)^x(1-p)^n = \binom{n}{x}\exp\Big(x\log\Big(\frac{1}{1-p}\Big) - (-n\log(1-p))\Big).$$

Canonical form of the exponential family:

$$f(x;\eta) = h(x)\ \exp(\eta^T T(x) - A(\eta)).$$

Suppose $T = \sum_{1\le i<j\le n} A_{ij}$. Let $X \in \{0,1\}^{\binom{n}{2}}$, then

$$\begin{aligned} f(x;\theta) &\propto \exp(\theta T) \\ &\propto \exp(\theta \sum_{1\le i<j\le n} a_{ij}) \end{aligned}$$

$$\begin{aligned} f(x;p) &= p^{\sum_{1\le i<j\le n} a_{ij}} (1-p)^{\binom{n}{2} - \sum_{1\le i<j\le n} a_{ij}} \\ &= \Big(\frac{p}{1-p}\Big)^{\sum_{1\le i<j\le n} a_{ij}} (1-p)^{\binom{n}{2}} \\ &= \exp T \log\Big(\frac{p}{1-p}\Big) - \binom{n}{2}\log(1-p) \\ &\propto \exp(T\theta). \end{aligned}$$

MATH 586 Fall 2023
Statistics for Networks

Lecture 17: Exponential Random Graph Models (Continued)

Lecturer: Robert Lunde Scribe: Yi Luo

17.33 Review

17.33.1 Exponential families

$P = \{P_\theta : \theta \in \Theta\}$ is an exponential family if the pdf/pmf permits a representation of the form:

$$f(x;\theta) = h(x)\exp(\eta(\theta)^\top T(x) - A(\theta)).$$

By factorization theorem, $T(x)$ is sufficient.

17.33.2 Canonical form of Exp family

Let $\eta = \eta(t)$. Consider parameterization:

$$f(x;\eta) = h(x)\exp(\eta^\top T(x) - A(\eta))A(\eta) = \log \int \frac{h(x)\exp(T^\top \eta)}{\text{log normalizer}} d\mu.$$

17.34 Exponential Random Graph Modes (ERGMs)

- Sufficient statistics $T(x) = (T_1(x), \cdots, T_k(x))$.
- Construct exponential family of form $f(x;\theta) = \exp(T(x)^\top\theta)/z(\theta)$, where $x \in \{0,1\}^{\binom{n}{2}}$.

17.34.1 MGD of Exponential Families

$$\begin{aligned}
M_T(s) &= Ee^{s^\top T} \\
&= \int \exp(s^\top T) h(x) \exp(\eta^\top T - A(\eta)) d\mu \\
&= \exp(-A(\eta)) \int \exp((s+\eta)^\top T(x)) h(x) d\mu(x) \\
&= \exp(A(\eta+s) - A(\eta)) \\
K_T(s) &= \log M_T(s).
\end{aligned}$$

For exponential families,

$$\begin{aligned}
K_T(s) &= A(\eta+s) - A(\eta) \\
\frac{\partial}{\partial s_i}[K_T(s)]_{s=0} &= \lim_{n\to 0} \frac{A(\eta+he_i) - A(\eta)}{h} \\
&= \frac{\partial}{\partial \eta_i} A(\eta) = ET_i \\
\frac{\partial^2}{\partial s_i \partial s_j}[K_T(s)]_{s=0} &= \mathrm{Cov}(T_i, T_j).
\end{aligned}$$

where $e_i =$ denotes the unit vector such that $(e_i)_i = 1$ and $(e_i)_j = 0$ for $j \neq i$.

Similarly,

$$\frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\eta) = \mathrm{Cov}(T_i, T_j).$$

Consider maximum likelihood equation of η,

$$\begin{aligned}
L(x;\eta) &= h(x) \exp(\eta^\top T(x) - A(\eta)) \\
l(x;\eta) &= \log L(x_j, \eta) - \log h(x) + \eta^\top T(x) - A(\eta).
\end{aligned}$$

Then,

$$\frac{\partial l}{\partial \eta_i} = 0 \Rightarrow T_i(x) - \frac{\partial}{\partial n_i} A(\eta) = 0 \Rightarrow T_i(x) = E_\eta T_i(x).$$

Expectations of Exponential (λ) is $1/\lambda$, expecation of Poisson (λ) is λ, matrix of mixed partials is given by $-\Sigma_T$.

$$A(\eta) = \log \int h(x) \exp(T^\top(x)\eta) d\mu(x).$$

For ERGMs, $x \in \{0,1\}^{\binom{n}{2}}$,

$$A(\eta) = \log \sum_{x \in X} h(x) \exp(T^\top(x)\eta).$$

For $h(x) = 1$,

$$\frac{\partial}{\partial \eta} A(\eta) = \frac{\sum_{x\in X} \exp(\eta^\top T(x)_i) T_i(x)}{\sum_{x \in X} \exp(\eta^\top T(x))}.$$

17.35 One alternative method for estimation: MCMCMLE

Consider ERGM with distribution of the form: $p(x;\theta) = \exp(T(x)^\top\theta)/Z(\theta)$.

Since

$$\frac{L(x;\theta)}{L(x;\theta_0)} = \frac{\exp(T^\top\theta)}{Z(\theta)} / \frac{\exp(T^\top\theta_0)}{Z(\theta_0)} = \exp(T^\top(\theta-\theta_0))/(Z(\theta)/Z(\theta_0)),$$

we can express MLE as:

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta_0} \frac{L(x;\theta)}{L(x;\theta_0)} = \arg\max_{\theta\in\Theta_0} \exp(T^\top(\theta-\theta_0))/(Z(\theta)/Z(\theta_0)).$$

This leads to

$$\sum_{x\in X} \exp(T(x)^\top\theta)/Z(\theta) = 1$$
$$\Rightarrow Z(\theta) = \sum_{x\in X} \exp(T^\top\theta),$$

and thus

$$\frac{Z(\theta)}{Z(\theta_0)} = \sum_{x\in X} \frac{e^{T(x)^\top\theta}}{Z(\theta_0)} = \sum_{x\in X} e^{T(x)^\top(\theta-\theta_0)} \frac{e^{T(x)^\top\theta_0}}{Z(\theta_0)} = E_{\theta_0}(e^{T(x)^\top(\theta-\theta_0)}).$$

We can approximate $E_{\theta_0}(e^{T(x)^\top(\theta-\theta_0)})$ via simulation

$$E_{\theta_0}(e^{T(x)^\top(\theta-\theta_0)}) \approx \frac{1}{B}\sum_{i=1}^{B} e^{T(Y^{(i)})^\top(\theta-\theta_0)},$$

where $Y^{(1)},\cdots,Y^{(B)}$ are simulated from $p(x;\theta_0)$.

We want Markov Chain with stationary measure $\pi(x) = p(x;\theta_0)$:

$$\rho(x,y) = \min(\frac{\pi(y)p(x|y)}{\pi(x)p(y|x)}, 1).$$

MATH 586 Fall 2023
Statistics for Networks

Lecture 18: Properties of ERGM Model's

Lecturer: Robert Lunde Scribe: Jingtao Shang

18.36 Review

- Exponential Family

$$P(x;\theta) = h(x)\exp(\eta(\theta)^T T(x) - A(\theta)$$

Canonical Form:

$$P(x;\eta) = h(x)\exp(\eta^T T(x) - A(\eta))$$

$T(x)$ is a sufficient statistic. (T is sufficient for θ if $X|T = t$ does not depend on θ)

- Exponential Random Graph Model

  Choose sufficient statistic $T = (T_1(x, \ldots, T_k(x))$
  Construct exponential family as form:

$$p(x;\theta) = \exp(T(x)^T\theta)/Z(\theta),\ where\ x \in \{0,1\}^{\binom{2}{n}}$$

  .

18.37 Problem with fitting ERGM

- Normalization factor $Z(\theta)$ is difficult to estimate.
  -One solution: MCMCMLE.
- Degeneracy
  Model puts disproportionate mass on a few possible configurations of the graph.
  For models like edge-triangle model, ERGM tends to put a lot of mass on complete/empty graphs.
- Estimation method (e.g MCMCMLE) don't often perform well
  Bhamidi, Bresler, Sly (2008) shows that if θ is non-negative, then one of the following are true with n large:
  -ERGM is essentially the same as Erdos-Renyi.
  or
  -ERGM mixes exponentially slowly.
- Chatterjee and Diaconis (2013) extend result above and show under condition $\theta \geq 0$, ERGM with large n behaves like ER or mixture of ER.
- Some proponents of ERGM's argue that asymptotic regime where the number of sufficient statistics fixed with n is unrealistic.
  -The number of sufficient statistics to change should be allowed.
  -However, sufficiency is a strong condition-questionable whether constructing models based on sufficient statistics would fit well to begin with.
- Instability
  Small changes in parameters lead to large changes in probabilistic behavior.

$$p(x;\theta) = \exp(T(x)^T\theta)/Z(\theta)$$

18.38 Pros and Cons of ERGM's

- Pros
  -It can construct relatively intuitive/ interpretable models.
  -Inference for parameters is well understood. (e.g. Asymptotic Normality of MLE)

- Cons
  -Models can be difficult to fit.
  -Common parametrizations/setups may be ill-behaved.
  -Sufficient statistics likely aren't sufficient for real-world graphs.

MATH 586 Fall 2023
Statistics for Networks

Lecture 19: Block Models

Lecturer: Robert Lunde Scribe: Jingtao Shang

19.39 Block Models

- Each vertex assumed to belong to one of k communities 1,...,k.
- Generative model
$$A_{ij} \sim Bernoulli(B_{rs})$$
,
where r is conmmunity of node i, s is the community of node j, B is $k \times k$ matrix of connection probability.

19.40 Stochastic Block Models

- It is often the case community assignments are unknown/random.
- Generative Model:
  -Let $\alpha = (\alpha_1, \ldots, \alpha_k)$ denote the vector where α_j is probability a node belongs to community j.
  -Let $Z_1, \ldots, Z_n \sim Multinomial(1, \alpha)$.
  -$A_{ij} \sim Bernoulli(B_{Z_i, Z_j})$ and $P(A_{ij}|Z_i = r, Z_j = s) = B_{rs}$
- Joint PMF of SBM
  Assuming $Z_1, \ldots, Z_n$ are known
$$L(B, \alpha) = f(Z, A) = \prod_{i=1}^{n} \alpha_{Z_i} \prod_{1 \le i < j \le n} B_{Z_i Z_j}^{A_{ij}} (1 - B_{Z_i Z_j})^{1 - A_{ij}}$$
  If $Z_1, \ldots, Z_n$ are unknown,
$$L(B, \alpha) = \sum_{Z \in \{1, \ldots, k\}^n} f(Z, \alpha)$$
$$P(A_{ij}) = E(P(A_{ij}|Z_j, Z_j)) = \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i B_{ij} \alpha_j$$
  So, what is the distribution of $D_i = \sum_{i \ne j} A_{ij}$?
$$P(D_i \le t) = E[P(D_i \le t|Z_i)], \ where \ D_i|Z_i = r \sim Binomial(n - 1, \sum_{j=1}^{k} B_{rj} \alpha_i)$$

- Low rank nature of SBM's -Let $P_{ij} = P(A_{ij} = 1|Z_i, Z_j)$
  -Suppose we rearrange labels so that communities are grouped together, for simplicity, suppose we have ordered them from 1 to k. Then,

$$P = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1k} \\ B_{21} & B_{22} & \dots & B_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ B_{k1} & B_{k2} & \dots & B_{kk} \end{bmatrix}$$

  We can also express P as

$$P = \Theta B \Theta^T$$

  where $\Theta = (\theta_1, \dots, \theta_n)^T \in \mathbf{R}^{n \times k}$ is a matrix such that each now θ_i contains only one non-zero entry, where the Θ_{ij} is equal to 1 only if node i belongs to community j.

$$\Theta B = \begin{bmatrix} B_{1r_1} & \dots & B_{1r_n} \\ \vdots & \ddots & \vdots \\ B_{nr_1} & \dots & B_{nr_n} \end{bmatrix}, \text{ where } r_1, \dots, r_n \text{ are memberships of nodes } 1, \dots, n$$

$$((\Theta B)\Theta^T)_{ij} = B_{r_j r_i} = B_{r_i r_j} \text{ (assuming the graph is undirected)}$$

- Mixed Membership SBM
  Suppose $\eta_1, \dots, \eta_n \overset{iid}{\sim} P$, where $\sum_{j=1}^k \eta_{ij} = 1$ and $A_{ij} \sim Bernoulli(\eta_i^T (B_{ij} \eta_j))$

- Degree-corrected SBM
  Let $\theta_1, \dots, \theta_n$ be additional degree parameters for each node (fixed or random). Then,

$$A_{ij} \sim Bernoulli(\theta_i \theta_j B_{Z_i Z_j}), \text{ where } Z_1, \dots, Z_n \sim Mutinomial(1, \alpha)$$

MATH 586 Fall 2023
Statistics for Networks

Lecture 20: Community Detection

Lecturer: Robert Lunde Scribe: Shourjo Chakraborty

20.41 Introduction

Goal: Suppose we have a sample of n network data points with k communities. Our goal is to accurately estimate the community labels, say $Z_1, Z_2, \dots, Z_n$ up to a permutation.

Example: Let $n = 3$ and $k = 2$. Suppose the orginal labels of X_1, X_2 is 1 and of X_3 is 2 whereas the estimated labels of X_1, X_2 is 2 and of X_3 is 1. In this case, the communities are detected correctly so we can change the estimated labels by applying a permutation π such that $\pi(1) = 2$ and $\pi(2) = 1$.

Notions of performance of community estimation/clustering:

- A natural statistic following the example above will be

$$T(z, \hat{z}) = \frac{1}{n} \max_{\pi \in S_k} \sum_{i=1}^{n} I(z_i = \pi(\hat{z}_i))$$

  We want $T(z, \hat{z})$ to be close to 1.

- It is also meaningful to consider

$$\tilde{T}(z, \hat{z}) = \min_i \max_{\pi \in S_k} \frac{\sum_{j=1}^{n} I(z_j = \pi(\hat{z}_j), z_j = i)}{\sum_{l=1}^{n} I(z_l = i)}$$

  Here, we are looking at the proportion of correctly estimated labels within each community and we want the minimum of these to be as large as possible which means $\tilde{T}(z, \hat{z})$ close to 1 is desired.

Notions of consistency: We are interested to look at the performances of T or $\tilde{T}$ or any other statistic in the asymptotic setup. For this we define the following two notions of consistency.

- Strong consistency/Exact recovery: $\mathbb{P}(T(z, \hat{z}) = 1) = 1) = 1 - o(1)$
- Weak consistency/Almost exact recovery: $\mathbb{P}(T(z, \hat{z}) = 1 - o(1)) = 1) = 1 - o(1)$

20.42 Community Detection algorithms

One of the most commonly used algorithm for community detection is spectral clustering.

Algorithm 2 Spectral clustering for community detection in network data

1: Compute $A = \hat{V}\hat{D}\hat{V}^\intercal$
2: Run clustering algorithm (k-means, k-medians, etc) on $\hat{V}_k \in \mathbb{R}^{n \times k}$
3: Return the clusters as estimated communities.

(*Note:* Recall that the matrix P (expectation of the adjacency matrix A conditioned on community labels) is low rank. So, $P = VDV^\intercal$ where $\lambda_i = 0$ for all $i > k + 1$.)

Proof: To build intuition, suppose $k = 2, n = 2m$, B is of the form $B = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$. Moreover, suppose $\{1, \ldots, m\}$ belong to class 1, $\{m+1, \ldots, 2m\}$ to class 2. So, P will be

$$P = \left(\begin{array}{ccc|ccc} p & \cdots & p & q & \cdots & q \\ \vdots & & \vdots & \vdots & & \vdots \\ p & \cdots & p & q & \cdots & q \\ \hline q & \cdots & q & p & \cdots & p \\ \vdots & & \vdots & \vdots & & \vdots \\ q & \cdots & q & p & \cdots & p \end{array}\right)$$

Each of the blocks in P have dimension $m \times m$. Clearly P is of rank 2. The eigenvalues of P are $\lambda_1 = \frac{n(p+q)}{2}$ and $\lambda_2 = \frac{n(p-q)}{2}$ with corresponding eigenvectors $v_1 = \frac{1}{\sqrt{n}}(1, 1, \ldots, 1)$ and $v_2 = \frac{1}{\sqrt{n}}(1, \ldots, 1, -1, \ldots, -1)$.

MATH 586 Fall 2023
Statistics for Networks

Lecture 21(a): Spectral Clustering

Lecturer: Robert Lunde Scribe: Ty Easley (Oct. 20, 2023)

21.43 Review (Community Detection in SBMs)

Stochastic Block Model

Definition 21.43.1 (SBM(α, B))**.** Suppose G is a graph with adjacency matrix $A \in M(n)$ has K communities, and node-wise G community memberships are described by i.i.d generalized Bernoulli variables $Z_1, \ldots, Z_n \sim P$, with $P(Z_i = k) = \alpha_k$. We have

- $A_{ij} \in \text{Bernoulli}(B_{Z_i Z_j})$
- $P = E[A \mid Z_1, \ldots, Z_n] = \Theta B \Theta^T$, where $\mathbb{R}^{n \times K} \ni \Theta = (\theta_1, \ldots, \theta_n)$ and $\Theta_{ij} = 1$ if and only if node i is a member of community j.

We call G the ***stochastic block model parametrized by*** (α, B) or SBM(α, B).

Community Detection via Spectral Clustering

To estimate community memberships $Z_1, \ldots, Z_n$, we can perform spectral clustering via the following algorithm:

1. Compute spectral decomposition of $A = VPV$
2. Take $V_k \in \mathbb{R}^{n \times k}$, corresponding to the largest k eigenvalues
3. Run clustering algorithm of your choice (e.g., k-means, k-medians) on V_k
4. Return cluster assignments as estimated community memberships

21.44 Spectral Clustering (an example)

Consider the community detection example from the previous lecture, where we have a graph G with

- a vertex set $|\mathcal{V}| = 2m$ with vertces $1, \ldots, m$ in community C_1 and vertices $m+1, \ldots, 2m$ in community C_2,
- an edge with probability p if it connects two vertices in the same community and probability q if it connects two vertices in different communities,

which we may summarize as

$$\begin{aligned} B &= \begin{bmatrix} p & q \\ q & p \end{bmatrix}, \\ \implies P &= \Theta B \Theta^T \\ &= \left[\begin{array}{c|c} p\mathbb{1}_m\mathbb{1}_m^T & q\mathbb{1}_m\mathbb{1}_m^T \\ \hline q\mathbb{1}_m\mathbb{1}_m^T & p\mathbb{1}_m\mathbb{1}_m^T \end{array}\right]. \end{aligned}$$

Spectral decomposition then gives us a 2-dimensional eigenspace $v = (v_1, v_2)$ spanned by

$$\begin{aligned} \lambda_1 &= n\left(\frac{p+q}{2}\right), & v_1 &= (1,\dots,1,1,\dots,1)/\sqrt{n} \\ \lambda_2 &= n\left(\frac{p-q}{2}\right), & v_2 &= (1,\dots,1,-1,\dots,-1)/\sqrt{n} \end{aligned}$$

Remark 21.44.1. Note that we assume $p > q$ here; if $p = q$, then this reduces to the Erdos-Renyi model, and if $p < q$, then we would simply swap labels between "in-" and "out-community" edges to recover the same model.

Bounding the approximation error

We now define the error term $E = A - P$, i.e., the error in estimating A and the ensuing community memberships: we wish to place a bound on the norm of E. Random matrix theory gives something of an all-purpose bound

$$\|E\|_{op} \le C\sqrt{n},$$

but it is, in general, difficult to make a sharp estimate of C. However, taking $p \to 0$ (i.e., $p_n \to 0$) at a given sparsity rate, we have a bound of order $\|E\|_{op} \le \sqrt{np_n}$. Define

$$\delta = \min\left(\lambda_2/n, (\lambda_1 - \lambda_2)/n\right).$$

If we take $\hat{v}_i$ to be an eigenvalue of P (i.e., an estimate of an eigenvalue v_i of A) and define $\theta(u, w) = \arccos(u \cdot w/\|v\|\|w\|)$, then a a variant of the Pauls-Kahan theorem yields

$$\begin{aligned} \sin\theta(\hat{v}_2, v_2) &\le \frac{\|E\|_{op}}{n\delta} \\ &\le \frac{\delta C}{\sqrt{n}} \\ &\to 0. \end{aligned}$$

Since $v_2 = (1,\dots,1,-1,\dots,-1)/\sqrt{n}$, we can write

$$\begin{aligned} \sin\theta(\hat{v}_2, v_2) &= \frac{1}{\sqrt{n}} \min_{s=\pm 1} \|s\sqrt{n}\hat{v}_2 - \sqrt{n}v_2\|_2 \\ \implies \min_{s=\pm 1} \frac{n}{2} \sum_{i=1}^{n} |e_i^T(s\hat{v}_2 - v_2)| &\le \frac{C^2}{\delta^2}, \end{aligned}$$

where e_i are the canonical basis vectors of $\mathbb{R}^n$ and the last inequality follows from the lemma below.

Lemma 21.44.2. *For unit vectors $\hat{v}_i, v_i$ and $\theta_i \coloneqq \arccos(\hat{v}_i^T v_i)$, we have*

$$\min_{s=\pm 1} \|s\hat{v}_i - v_i\|_2 \le \sqrt{2}\sin\theta_i.$$

Proof. Choose s so that $s\hat{v}v_i^T v_i \geq 0$. Let $\tilde{v}_i = s\hat{v}_i$ and write

$$\begin{aligned}\min_{s=\pm 1} \|s\hat{v}_i - v_i\|_2 &\leq \|\tilde{v}_2 - v_2\|_2 \\ &\leq \sqrt{2}\left(1 - (\tilde{v}_i^T v_i)^2\right)^{1/2} \\ &= \sqrt{2}\sin\theta_i\end{aligned}$$

□

21.45 Eigenstructure of SBM/Degree-Corrected SBM

In the example discussed above, we used the eigenstructure of P to be bound the error of our approximation; to generalize this approach, we need to broaden our understanding of the eigenstructure of P in the SBM case.

Theorem 21.45.1. *Suppose $B \in \mathrm{GL}(K, \mathbb{R}) \subset M(K, \mathbb{R})$ is full rank and $P = \Theta B \Theta^T$. Then P has an eigendecomposition $P = UDU^T$ where $U = \Theta X$ and $X \in M(K, \mathbb{R})$.*

Proof. Let $\Delta = \mathrm{diag}(\sqrt{n_1}, \ldots, \sqrt{n_K})$, where n_j is the number of vertices belonging to community j. Since $\Theta\Delta^{-1}$ is orthonormal, we write P as

$$\begin{aligned}P &= \Theta B \Theta^T \\ &= \Theta\Delta^{-1}\left(\Delta B \Delta\right)\left(\Delta^{-1}\right)^T \\ &= UDU^T,\end{aligned}$$

where the eigendecomposition $B = VDV^T$ gives $U = \Theta\Delta^{-1}V$. Then $X = \Delta^{-1}V$ is orthornormal. □

We also wish to consider the case of the degree-corrected SBM; here, we recover an analogous result, with the (similar) main takeaway that we decompose P into degree-dependent and degree-independent terms.

Theorem 21.45.2. *Let Ψ be the vector of degree parameters corresponding the degree-corrected SBM defined by P, which we suppose has the form*

$$UDU^T = P = \mathrm{diag}(\Psi)\Theta B \Theta^T \mathrm{diag}(\Psi).$$

Then there exists $H \in \mathrm{SO}(K)$ such that

$$U_i = \widetilde{\Psi}_i H,$$

where $\widetilde{\Psi}_i$ depends only on the degree parameter of node i.

21.46 Review

SBM(α, β)
-SBM is a random graph model, which tends to produce graphs containing communities and assigns a probability value to each pair i, j (edge) in the network.
- To perform community detection, one can fit the model to observed network data using a maximum likelihood method.

- Suppose there K communities
- $Z_1, \ldots, Z_n \sim P$, where $P(Z_i = k) = \alpha_k$, $A_{ij} \sim \text{Bernoulli}\left(B_{Z_i Z_j}\right)$

$$P = \mathbb{E}[A \mid Z_1, \cdots, Z_n] \\ = \Theta B \Theta^\top \quad ,$$

where $\Theta \in \mathbb{R}^{n \times k}$ and $\Theta = (\theta_1, \cdots, \theta_n)$, $\theta_{ij} = \begin{cases} 1, \text{if } i \text{ belongs to community } j, \\ 0, \text{otherwise} \end{cases}$

21.47 Spectral Clustering

Community Detection:
Goal: Estimate community membership: $Z_1, \cdots, Z_n$.

21.47.1 Algorithm

Spectral Clustering:
1. Compute spectral decomposition of $A = \hat{V} D \hat{V}^\top$,
2. Take $\hat{V}_k \in \mathbb{R}^{n \times k}$. corresponding to k eigenvalues with largest magnitude,
3. Rum clustering algorithm (e.g. k means, k medians),
4. Retum cluster assignments as estimated communities.

21.47.2 Example

Suppose $1, \cdots, m$ belong to community 1,
$m+1, \cdots, 2m$ belong to en community 2, then we have:
$B = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$, $P = \Theta B \Theta^\top = \begin{bmatrix} p\mathbb{1}_m \mathbb{1}_m^\top & q\mathbb{1}_m \mathbb{1}_m^\top \\ q\mathbb{1}_m \mathbb{1}_m^\top & p\mathbb{1}_m \mathbb{1}_m^\top \end{bmatrix}$, $n = 2m$

$\lambda_1 = \left(\frac{p+q}{2}\right) \times n, \quad v_1 = (1, \ldots 1)/\sqrt{n}$
$\lambda_2 = \left(\frac{p-q}{2}\right) \times n \quad v_2 = (\underbrace{1, \ldots, 1}_{m}, \underbrace{-1, \ldots, -1}_{m})$

The expression is $A = P + E$, with the signal noise: $E = A - P$.

We need to know magnitude of $\|E\|_{\text{op}}$.
- A general purpose bound from random matrix theory gives:

$$\|E\|_{op} \leqslant c\sqrt{n}, \text{ w.h.p.}$$

(it's not tight and can be sharper considering the value of p, q)

Essentially, if $p \to 0$, for a certain sparsity range, we have bound of order: $\|E\|_{op} \leq \sqrt{np_n}$.
The noise has lower order than the signal(this is the intuition of why the method works).
$\|p\|_{\text{lop}} \sim \left(\frac{p+q}{2}\right) \times n$
$\|E\|_{\text{op}} \leq c\sqrt{n}$ w.h.p.
Let $\delta = \min\left(\frac{\lambda_2}{n}, \frac{\lambda_1 - \lambda_2}{n}\right)$, a variant of Davis-Kahan yields:

$$\sin\Theta\left(\hat{v}_2, v_2\right) \leq \frac{\|E\|_{\text{op}}}{n \times \delta}$$
$$\leq \frac{c\sqrt{n}}{n\delta} \leq \frac{c}{\sqrt{n}\delta}.$$

$\sqrt{n}v_2 = \{1, \ldots 1, -1, \ldots -1\}$
$\sqrt{n}\sin\theta(\hat{v}_2, v_2)\sqrt{\sum_{i=1}^{n}\left(\hat{v}_{2,i} - 1\right)^2}$
$= \min_{c\in\{-1,1\}} \|c\sqrt{n}\hat{v}_2 - \sqrt{n}v_2\|.$
It follows that:
$\#|\left\{i \in \{1, 2, \cdots, n\}, \text{sign}\left(c\hat{v}_2 \neq v_2\right)\right\}| \leq \frac{c^2}{\delta^2}$

21.48 Eigenstructure of SBM

Suppose $B \in \mathbb{R}^{k\times k}$ full rank, $P = \Theta B\Theta^\top$. Then $P = UDU^\top$, where $U = \Theta X, X \in \mathbb{R}^{\text{k}\times\text{k}}$
Proof:
Let $\Delta = \text{diag}\left(\sqrt{n_1}, \cdots, \sqrt{n_k}\right)$, where n_j is # vertices belong to community j.
Express P as:
$$P = \left(\theta\Delta^{-1}\right)\Delta B\Delta\left(\theta\Delta^{-1}\right)^\top$$
Observation:
$$\left(\theta\Delta^{-1}\right)^\top\left(\theta\Delta^{-1}\right) = I_k$$
Let $ZDZ^\top$ be eigen-decomposition of $\triangle B\Delta$, we can show: $\theta\Delta^{-1}z$ are orthonormal, with
$$\left(\theta\Delta^{-1}z\right)^\top\left(\theta\Delta^{-1}z\right)$$
$$z^\top\left(\theta\Delta^{-1}\right)^\top\left(\theta\Delta^{-1}z\right)$$

MATH 586	Fall 2023
Statistics for Networks	
Lecture 22:Community Detection	
Lecturer: Robert Lunde	Scribe: Zongxi Yu

22.49 Review

SBM(β, α)
- Suppose there are k communities.
- $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} P$, where $P\left(Z_i = j\right) = \alpha_j$.
- $A_{ij} = A_{ji} \sim \text{Bernoulli}\left(B_{Z_iZ_j}\right)$.
- Let $P = \mathbb{E}\left[A \mid Z_1, \cdots, Z_n\right]$.
- $P = \Theta B\Theta^\top$ where $\Theta \in \mathbb{R}^{n\times k}$, and $\theta_{ij} = \begin{cases} 1, \text{if node } i \text{ belongs to community } j \\ 0, \text{otherwise} \end{cases}$.

22.50 Lemma

Representation of P matrix
Consider eigen-decomposition $P = UDU^\top$. Then, $U = \Theta X, X \in \mathbb{R}^{k \times k}$. (assuming rank of B is k)

Lemma for DCSBM
Consider mean matrix $P = diag(\psi)\Theta B \Theta^\top diag(\psi)$, where $\psi = (\psi_1, \psi_2, \cdots, \psi_n)$ are degree parameters. Let $P = UDU^\top$, where $U_{ik} = \psi_i H_{k*}$(belongs to community$k$, $\hat{\psi}_2$ is related to degree parameter).

Recall example where $B = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$ nodes $1, \cdots, m$ belong to community 1, nodes $m+1, \cdots, 2m$ belong to community 2.

Claim: For $\hat{v_i}$,v_i unit vectors, $\min_{c \in \{-1,1\}} \|c\hat{v_i} - v_i\|_2 \leq \sqrt{2} \sin\theta_i$.

Proof
Choose c so that $\hat{v_i}^\top v_i \geq 0$: let $\hat{v_i} = cv_i$ for this choice of c.

$$\begin{aligned}
\min_{c \in \{-1,1\}} \|c\hat{v_i} - v_i\|_2 &\leq \|\hat{v_i} - v_i\|_2 \\
&= \sqrt{\sum_{j=1}^{d} (\hat{v}_{ij} - v_{ij})^2} \\
&= \sqrt{\sum_{j=1}^{d} (\hat{v}_{ij}^2 + v_{ij}^2 - 2(\hat{v}_{ij} v_{ij}))} \\
&= \sqrt{2(1 - \hat{v_i}^\top v_i))} \\
&= \frac{\sqrt{2}\sqrt{(1 - \hat{v_i}^\top v_i)(1 + \hat{v_i}^\top v_i)}}{\sqrt{(1 + \hat{v_i}^\top v_i)}} \\
&\leq \sqrt{2}\sqrt{(1 - \hat{v_i}^\top v_i)^2} \\
&\leq \sqrt{1 - \cos^2\theta_1} (\text{Recall: } \sin^2\theta + \cos^2\theta = 1) \\
&\leq \sqrt{2}\sqrt{\sin^2\theta_i}
\end{aligned}$$

Support that we are clustering based on sign of $\hat{v}_2$, we misclassify node i if $(\sqrt{n}\hat{v}_2(i) - \sqrt{n}v_2(i) \geq 1$.

$$\begin{aligned}
\sum_{i=1}^{n} \frac{1}{n} \mathbf{1}_{\text{node i misclassified}} &\leq \min_{c \in \{-1,1\}} \frac{1}{n} \|\sqrt{n}(\hat{v}_2 - v_2)\|_2^2 \\
&= \frac{1}{n} \sum_{j=1}^{d} (\sqrt{n}(\hat{v}_2(j) - v_2(j)))^2
\end{aligned}$$

Strong consistency typically holds when $\frac{np_n}{\lg n} \to \infty$, weak consistency when $np_n \to \infty$.

22.51 Two Spectral Procedures for DCSBM

SCORE
1. Compute eigen-decomposition $A = \hat{v}\hat{D}\hat{v}^\top$.
2. Construct matrix $R \in \mathbb{R}^{n\times(k-1)}$, where $R_{ij} = \frac{\hat{v}_{j+1}(i)}{\hat{v}_1(i)}$.
3. Run clustering algorithm (e.g k-means, k-medians on R).
4. Return k clusters.

Spherical Spectral Clustering
1. Compute eigen-decomposition $A = \hat{v}\hat{D}\hat{v}^\top$.
2. Normalize each row of $\hat{v} \in \mathbb{R}^{n\times k}$ by its norm. Let v_k^* be resulting matrix.
3. Run k-means/k-median clustering on v_k^*.
4. Return k clusters.

22.52 Likelihood-based Methods

Karner and Newman proposed Poisson profile likelihood.
$A_{ij} = A_{ji} \sim \text{Poisson}\left(\lambda_{Z_iZ_j}\right)$, where $\lambda_{Z_iZ_j} = B_{Z_iZ_j}$.

They also assume $A_{ii} = 2 \times \mathbf{1}_{i \text{ has self loop}}$.

$$P(\mathcal{G}|\lambda, z) = \prod_{1\le i<j\le h} \frac{(\lambda_{Z_iZ_j})^{A_{ij}} exp(-\lambda_{Z_iZ_j})}{(A_{ij})\,!} \times \prod_{i=1}^{n} \frac{\frac{1}{2}(\lambda_{Z_iZ_i})^{A_{ii}/2} exp(-\frac{1}{2}\lambda_{Z_iZ_i})}{(\lambda_{ii}/2)\,!}$$

$$= (\lambda_{rs})^{ors/2} exp(-\frac{1}{2}O_rO_s\lambda_{rs}) \times \frac{1}{\prod_{1\le<i<j\le n} A_{ij}\,!} \prod_{i=1}^{n} 2^{A_{ij}/2}(A_{ij}/2)$$

where $ORS_i = \sum_{ij} A_{ij}(z_i = r, z_j = s)$

MATH 586 Fall 2023
Statistics for Networks

Lecture 23: Community Detection

Lecturer: Robert Lunde Scribe: Bahram Yaghooti

23.53 Review

23.53.1 SBM(α, β)

$z_1, \ldots, z_n \sim P$,
where $P(z_i = k) = \alpha_k$
$A_{ij} = A_{ji} \sim \text{Bernoulli}(B_{z_iz_j})$, and $B \in [0,1]^{k\times k}$.

23.53.2 Degree-corrected SBM

$z_1, \ldots, z_n \sim P$ as before
$\psi_1, \ldots, \psi_n$ degree parameters
$A_{ij} = A_{ji} \sim \text{Bernoulli}(\psi_i \psi_j B_{z_i z_j})$.

R is $\mathbb{R}^{n \times (k-1)}$ matrix
where $R_{ij} = \frac{V_{j+1}(i)}{V_1(i)}$

23.53.3 Community detection for degree-corrected SBM's

1. Score
   Run k-means/k-medians on eigenvector ratios
2. Spherical Spectral Clustering
   Run k-means/k-medians on normalized rows $\tilde{V}_k$.

23.54 Likelihood-Based Methods

Assume

$$A_{ij} = A_{ji} \sim \text{Poisson}(\lambda_{z_i z_j})$$
$$A_{ii} = 2 \times \mathbb{1}(i \text{ has self loop})$$

$$P(G|\lambda, z) = \frac{\prod_{i<j} \left(\lambda_{z_i z_j}\right)^{A_{ij}} \exp(-\lambda_{z_i z_j})}{\prod_{i=1}^n \frac{1}{2} (\lambda_{z_i z_j})^{A_{ii}/2} \exp\left(-\frac{1}{2}\lambda_{z_i z_j}\right)}$$

We can re-express as

$$\frac{1}{\prod_{i<j} A_{ij}! \prod_{i=1}^n 2^{A_{ii}/2} \left(\frac{A_{ii}}{2}\right)!} \times (\lambda_{rs})^{O_{rs}/2} \exp\left(-\frac{1}{2} n_r n_s \lambda_{rs}\right),$$

where $n_r = \#$ vertices in community r and O_{rs} is defined as

$$O_{rs} = \sum_{ij} A_{ij} \mathbb{1}(z_i = r, z_j = s)$$

Ignoring constants, Log-likelihood is of the form:

$$\log P(G|\lambda, z) = \sum_{rs} \left(O_{rs} \log \lambda_{rs} - n_r n_s \lambda_{rs}\right)$$

MLE is given by

$$\hat{\lambda}_{rs} = \frac{O_{rs}}{n_r n_s}$$

Plugging this into profile likelihood, ignoring constants, likelihood based criteria for community detection

$$\log P(G|z) = \sum_{rs} O_{rs} \log \frac{O_{rs}}{n_r n_s}$$

Pros of likelihood-based method

- Consistent under mild condition

Con: Computationally expensive (need to compute quantify for each possible assignment of latent positions to find maximum).

For DCSBM, a similar derivation yields

$$\mathcal{L}(G|z) = \sum_{rs} O_{rs} \log \frac{O_r}{O_r O_s}$$

where $O_r = \sum_s O_{rs}$ and $O_{rs} = \sum_{ij} A_{ij} \mathbb{1}(z_i = r, z_j = s)$.

23.55 Modularity

Very popular method where one only considers edges with communities Modularity function:

$$Q(z) = \sum ij(A_{ij} - P_{ij})\mathbb{1}(z_i = z_j)$$

for a choice of null model for P.

- Common choices of null model: Erdös-Reyni, degree-corrected Erdos-Reyni.
- Goal: Find community membership that maximizes Q (surprise factor relative to null model with no communities).

Under ER model, P can be estimated as $P_{ij} = \frac{L}{n^2}$, where $L = \#$ edges and

$$Q_{ER}(z) = \sum_r \left(O_{rr} - \frac{n_r^2}{r^2} L \right)$$

For Newman-Girvan modularity, estimate of P_{ij} is $\frac{d_i d_j}{2}$

$$Q_{NM}(z) = \sum_r \left(O_{rr} - \frac{O_r^2}{L^2} \right)$$

- Newman-Girvan modularity only works when within-community probabilities are larger than some quantity related to between-community probabilities.
- There are computational approximations for NG.

- Typically for weak consistency, we need $np_n \to \infty$
- However certain methods have also been studied/shown to be consistent for $np_n \to c$
- Regularize $L = D^{-1/2} A D^{-1/2}$ for special clustering in very sparse graphs.

MATH 586 Fall 2023
Statistics for Networks

Lecture 24: Identifiability issues with RDPG

Lecturer: Robert Lunde Scribe: Giacomo Vedovati

24.56 Review

Graphons

$A_{ij} = A_{ji} \sim \text{Bernoulli}(W(\xi_i, \xi_j))$, where $\xi_1, \ldots, \xi_n \sim \text{Uniform}[0,1]$, unknown latent positions.

Consider the integral

$$\text{T}\, f(x) = \int_0^1 W(x,y) f(y) dy$$

Let $(\lambda_r)_{r\in\mathbb{N}}$, $(\varphi)_{r\in\mathbb{N}}$ be the associated eigenvalues and eigenvectors. Suppose $W(u,v)$ admits a finite rank representation of the form:

$$W(u,v) = \sum_{r=1}^{d} \lambda_r \phi_r(u) \phi_r(v)$$

We can rewrite $W(\xi_i, \xi_j)$ as $< x_i, x_j > - < y_i, y_j >$m where $x_{ik} = \sqrt{\lambda_k} v_k(i)$, $y_{il} = \sqrt{\lambda_l} v_l(i)$, with x corresponding to the positive and y to the negative eigenvalues.

RDPG

$A_{ij} = A_{ji} \sim \text{Bernoulli}(< x_i, x_j >)$ with $x_i \in \mathbb{R}^p$ and $x_1, \ldots, x_n$ i.i.d.

GRDPG

$A_{ij} = A_{ji} \sim \text{Bernoulli}(< x_i, x_j > - < y_i, y_j >)$ with $x_i, y_i \in \mathbb{R}^p$

We showed that SBM, MMSBM and PCSBM are all GRDPG.

24.57 Identifiability issues with RDPG

Let Q be an orthonormal matrix s.t. $Q^\top Q = I$, $QQ^\top = I$. Then:

$$\begin{aligned} < Qx_i, Qx_j > &= (Qx_i)^\top (Qx_j) \\ &= x_i^\top Q^\top Q x_j \\ &= x_i^\top x_j \\ &=< x_i, x_j > \end{aligned}$$

In other words, Qx and x give us some distribution for A. For generalized RDPG, if one assumes x and y uncorrelated, then the source of non identifiability are of the form $Q = (Q^x, Q^y)$. In fact

$$< Q^x x_i, Q^x x_j > - < Q^y y_i, Q^y y_j >=< x_i, x_j > - < y_i, y_j >$$

One way to circumvent the identifiability issues is to consider the internal parameters that don't depend on A.

For triangles, a natural notion of subgraph frequecy is given by:

$$T = \frac{1}{\binom{n}{3}} \sum_{1\le i<j<k<n} A_{ij} A_{jk} A_{ki}$$

For a given tuple (i,j,k), $A_{ij}A_{jk}A_{ki}$ is Bernoulli with mean

$$\theta = \mathbb{E}[(< x_i, x_j > - < y_i, y_j >)(< x_j, x_k > - < y_j, y_k >)(< x_k, x_i > - < y_k, y_i >)]$$

This is the same regardless the orientation of (X, y).

Sparse generalized RDPG models

$A_{ij} = A_{ji} \sim \text{Bernoulli}(P_n[< x_i, x_j >< y_i, y_j >])$ where $P_n \to 0$.

If we let $\hat{x}_i = \sqrt{P}_n x_i$, $\hat{y}_i = \sqrt{P}_n y_i$, then $P_n[< x_i, x_j >< y_i, y_j >] =< \hat{x}_i, \hat{x}_j > - < \hat{x}_i, \hat{x}_j >$

Estimating $z_i = (x_i, y_i)$

We can estimate $z_i = (x_i, y_i)$ with the adjacency spectral embedding. Take the eigendecomposition $A = VDV^\top$, where D is sorted so that the positive eigenvalues are listed first in descending order.

$$(\hat{x}_{i1}, \ldots, \hat{x}_{in}) = (\sqrt{\lambda_1} v_1(i), \ldots, \sqrt{\lambda_p} v_p(i))$$

$$(\hat{y}_{i1}, \ldots, \hat{y}_{in}) = (\sqrt{\lambda_{p+1}} v_{p+1}(i), \ldots, \sqrt{\lambda_q} v_q(i))$$

with p larger positive and q the smallest negative eigenvalues.

Suppose

$$p = \sum_{r=1}^{d} \lambda_r v_r v_r^\top$$

Then

$$< \hat{x}_i, \hat{x}_j > - < \hat{x}_i, \hat{x}_j > \approx P_{i,j} = \sum_{r=1}^{d} \lambda_r v_r(i) v_r(j)^\top$$

Theorem 24.57.1 (Error Bound for estimating $z_r = (x_i, y_i)$)**.** *Suppose $\hat{z}_1, \ldots, \hat{z}_n$ estimated by adjacency spectra embedding. Then $\exists c > 1$ such that if $p(n) = W(\frac{\log^{4c} n}{n})$*

There exist a sequence of transformations such that:

$$\max_{1 \le i < n} ||\hat{z}_i - z_i|| = O_p(\frac{\log n}{\sqrt{n}})$$

MATH 586	Fall 2023
Statistics for Networks	
Lecture 25	
Lecturer: Robert Lunde	Scribe: Tianwei Zhou

25.58 Review

RDPG Model

$A_{ji} = A_{ij} \sim Bernoulli(\langle x_i, x_j \rangle)$, where $x_1, \ldots, x_n \overset{\text{i.i.d}}{\sim} P$, $x_i \in \mathbb{R}^d$.

GRDPG Model

$A_{ji} = A_{ij} \sim Bernoulli\big(\langle x_i, x_j\rangle - \langle y_i, y_j\rangle\big)$, where $(x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d}}{\sim} P$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}^q$.

Error Bound for estimating $z_i = (x_i, y_i)$

Let $z_i = (x_i, y_i)$, $\tilde{Z}_i = \sqrt{\phi}_n z_i$.

Proposition 25.58.1. *Suppose $A^{(n)} \in \{0,1\}^n \prod^n$ generated by GPPG, with latent positions $\tilde{z}_1, \ldots, \tilde{z}_n$. Let $\hat{z}_1, \ldots, \hat{z}_n$ be adjacency spectra embedding. Then for $c > 1$ such that if $p(n) = W(\frac{\log^{4c} n}{n})$ and some sequence $(Q_n)_{n\geq 1}$*

$$\max_{1\leq i<n} ||\hat{z}_i - z_i|| = O_p(\frac{\log n}{\sqrt{n}}), \tag{25.1}$$

where Q_n are a sequence of transformation that involve orthogonal matrices acting on positive parts separately

$$Q_n = \begin{bmatrix} Q^{\oplus} & 0 \\ 0 & Q^{\ominus} \end{bmatrix}.$$

Theorem 25.58.2. *Under the same conditions as the previous proposition,*

$$\sqrt{n} W\big(Q_n \hat{z}_i - \tilde{z}_i\big) \xrightarrow{d} \mathbb{N}\big(0, \Sigma(z_i)\big), \tag{25.2}$$

where $\Sigma(x)$ is some covariance matrix that depends on unscaled latent position.

25.59 Exchangebility

Recall graphon model

$$A_{ji} = A_{ij} \sim Bernoulli\big(\omega(\xi_i, \xi_j)\big), \tag{25.3}$$

where $\xi_1, \ldots, \xi_n \overset{\text{i.i.d}}{\sim} \mathbb{U}[0,1]$, $\omega : [0,1]^2 \to [0,1]$ symmetric.

Definition 25.59.1. A distribution on a graph G is vertex exchangeable if for any permutation of node labels, we have the same distribution for the graph.

Exchangeability for vectors

Definition 25.59.2. The vector $(x_1, \ldots, x_n)$ is exchangeable if for any permutation $\sigma : [n] \mapsto [n]$,

$$(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) \overset{d}{=} (x_1, \ldots, x_n). \tag{25.4}$$

Observation:

- IID $X_1, \ldots, X_n$ are exchangeable. For simplicity, assume X_i has desity $f_X(x)$.

$$f_{x_1,\ldots,x_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i), \tag{25.5}$$

  which means exchangebility is weaker, or sometime a lot weaker than i.i.d.

- Exchangebility $\Rightarrow$ identical distributed since $\mathbb{P}(X_{\sigma(i)} \in A) = \mathbb{P}(X_i \in A)$ must hold $\forall\, \sigma$.
- They must have the same dependence structure accross all possible subsets. E.g

$$\begin{aligned}\mathrm{Cov}(X_{\sigma(i)}, X_{\sigma(j)}) &= \mathbb{E}X_{\sigma(i)}X_{\sigma(j)} - \mathbb{E}X_{\sigma(i)}\mathbb{E}X_{\sigma(j)} \\ &= \mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j = \mathrm{Cov}(X_i, X_j).\end{aligned}$$

Suppose $\big(X_i\big)_{i\geq 1}$ is an exchangeable sequence. That is for any $\sigma : \mathbb{N} \mapsto \mathbb{N}$,

$$\big(X_{\sigma(i)}\big)_{i\geq 1} \stackrel{d}{=} \big(X_i\big)_{i\geq 1}.$$

Theorem 25.59.3 (DeFinett's Theorem)**.** *Suppose we have an exchangeable sequence $\big(X_i\big)_{i\geq 1}$. Then, we can express $\big(X_i\big)_{i\geq 1}$ as:*

- *For some $H \sim \gamma$ (H works like a mixture component), $X_1, X_2 \ldots | H \overset{i.i.d}{\sim} P_H$.*
- *Alternatively, there exists Borel measurable g and U, $\big(U_i\big)_{i\geq 1}$, where $U, U_i \overset{i.i.d}{\sim} \mathbb{U}[0,1]$ such that*

$$\big(X_i\big)_{i\geq 1} \stackrel{d}{=} \big(g(U, U_i)\big)_{i\geq 1}. \tag{25.6}$$

- *Alternatively,*

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots) = \int \prod_{i=1}^{n} \theta(A_i) v\big(d(\theta)\big). \tag{25.7}$$

MATH 586 Fall 2023
Statistics for Networks

Lecture 26: Graphon

Lecturer: Robert Lunde Scribe: Zhichen Xu

Review

Exchangeablity A vector $(X_1, ..., X_n)$ is exchangeable if for any $\sigma : [n] \mapsto [n]$

$$(X_{\sigma(1)}, ..., X_{\sigma(n)}) \stackrel{d}{=} (X_1, ..., X_n)$$

- Intuition: future is like past

- Exchangeable $\Rightarrow$ identical marginals, identical dependencies (eg $\mathrm{cov}(X_{\sigma(i)}, X_{\sigma(j)}) = \mathrm{cov}(X_1, X_2)$)

- IID special case of exchangeable vectors

Suppose $(X_i)_{i\geq 1}$ is an exchangeable sequence, for all $\sigma : N \mapsto N$

$$(X_{\sigma(i)})_{i\geq 1} \stackrel{d}{=} (X_i)_{i\geq 1}$$

Theorem De Finetti's Theorem

The exchangeable sequence $(X_i)_{i\geq 1}$ admits the representation:

$$(X_i)_{i\geq 1} \stackrel{d}{=} (g(U, U_i))_{i\geq 1}$$

$U_i \overset{\text{iid}}{\sim} Uniform[0,1], U \sim Uniform[0,1]$, g is Borel measurable.

Equivalent to:

$$P(X_1 \in A_1, X_2 \in A_2, ...) = \int \prod_{i=1}^{\infty} \theta(A_i) v(d\theta)$$

Note: Not all exchangeable vectors can be represented as mixture of IID random variables.

Ex: Consider the following distribution:

$$P(X_1 = 1, X_2 = 0) = P(X_1 = 0, X_2 = 1) = \frac{1}{2}$$

$$P(X_1 = 1, X_2 = 1) = P(X_1 = 0, X_2 = 0) = 0$$

If can be represented,

$$0 = P(X_1 = 1, X_2 = 1) = \int p^2 \mu(dp)$$

$$0 = P(X_1 = 0, X_2 = 0) = \int (1-p)^2 \mu(dp)$$

Therefore, $p, (1-p)$ both equals to 0 $a.s.$ Contradiction occurs.

While finite exchangeable sequence don't always admit representation as mixture of IID distributions, they can often be approximated by such a distribution.

Def N extendability

A vector $(X_1, ..., X_k)$ is N-extendable if there exists a vector $(\hat{X_1}, ..., \hat{X_N}), N > k$, such that

$$(X_1, ..., X_k) \stackrel{d}{=} (\hat{X_1}, ..., \hat{X_k})$$

Ex: $X_1, ..., X_k \overset{\text{iid}}{\sim} N(0,1)$ Then it is extendable to $X_1, ..., X_n \overset{\text{iid}}{\sim} N(0,1)$

Theorem Suppose $(X_1, ..., X_k)$ is exchangeable, n-extendable and X_i discrete, with c possible values. Then exists a measure P^* that is a mixture of iid distributions such that:

$$\|P(X_1, ..., X_k) - P^*(X_1, ..., X_k)\|_{TV} \leqslant \frac{2ck}{n}$$

proof We start by characterizing $P_n = P(X_1, ..., X_n)$. If we condition on order statistics $X_{(1)} = x_1, ..., X_{(n)} = x_n$, then each permutation of $X_1, ..., X_n$ are equal likely. Thus,

$$\begin{aligned} P(X_1, ..., X_n) &= \sum P(X_1, ..., X_n) P(X_{(1)} = x_1, ..., X_{(n)} = x_n) \\ &= \sum W_u H_{un}(X_1, ..., X_n) \end{aligned}$$

where $H_{un}(X_1, ..., X_n)$ is pmf associated with drawing n balls without replacement from $X_{(1)}, ..., X_{(n)}$.

Now consider $P_k = P(X_1, ..., X_k)$. Then

$$\begin{aligned} P(X_1, ..., X_k) &= \sum_{(X_1,...,X_n)|X_1=x_1,...,X_k=x_k} \sum W_u H_{uk}(X_1, ..., X_n) \\ &= \sum W_u H_{uk}(X_1, ..., X_k) \end{aligned}$$

Now choose $P^*(X_1, ..., X_k) = \sum_u W_u M_{uk}(X_1, ..., X_k)$, where M_{uk} corresponds to drawing n balls with replacement from $X_{(1)}, ..., X_{(n)}$. Then,

$$\begin{aligned} \|P_k - P^*\| &= \|\sum W_u H_{uk} - \sum W_u M_{uk}\|_{TV} \\ &\leqslant \sum_u W_u \|H_{uk} - M_{uk}\|_{TV} \\ &\leqslant \frac{2ck}{n} \end{aligned}$$

MATH 586 Fall 2023
Statistics for Networks

Lecture 27: Graphon

Lecturer: Robert Lunde Scribe: Zhichen Xu

Review

Exchangeablity A vector $(X_1, ..., X_n)$ is exchangeable if for all permutations $\sigma : [n] \mapsto [n]$

$$(X_{\sigma(1)}, ..., X_{\sigma(n)}) \stackrel{d}{=} (X_1, ..., X_n)$$

A sequence $(X_i)_{i\geq 1}$ is exchangeable if for all bijections $\sigma : N \mapsto N$

$$(X_{\sigma(i)})_{i\geq 1} \stackrel{d}{=} (X_i)_{i\geq 1}$$

Theorem Di Finetti's Theorem

The exchangeable sequence $(X_i)_{i\geq 1}$ admits the representation:

$$(X_i)_{i\geq 1} \stackrel{d}{=} (g(U, U_i))_{i\geq 1}$$

$U_i \overset{\text{iid}}{\sim} Uniform[0, 1], U \sim Uniform[0, 1]$, g is Borel measurable.

- Finite exchangeable vectors may not have such a representation.

- However, if finite vector is part of larger exchangeable vector, it will be close to a mixture of IID random variables.

Definition Jointly exchangeable array

The array $(A_{ij})_{1\leqslant i,j\leqslant n}$ is jointly or vertex exchangeable for any permutation $\sigma : [n] \mapsto [n]$ we have:

$$(A_{\sigma(i)\sigma(j)})_{1\leqslant i,j\leqslant n} \stackrel{d}{=} (A_{ij})_{1\leqslant i,j\leqslant n}$$

Similar to Di Finetti's theorem, in order to state representation theorems, we consider an infinite exchangeable array.

Definition Exchangeable infinite array

$(A_{ij})_{i,j\in N}$ is jointly exchangeable if for any bijection $\sigma : N \mapsto N$, we have

$$(A_{\sigma(i)\sigma(j)})_{i,j\in N} \stackrel{d}{=} (A_{ij})_{i,j\in N}$$

Theorem Aldous-Hoover Theorem

If $(A_{ij})_{i,j\in N}$ is jointly exchangeable, then there exists a Borel measurable f such that:

$$(A_{ij})_{i,j\in N} \stackrel{d}{=} (f(U, U_i, U_j, U_{\{\{i,j\}\}})$$

where $U, U_i, U_j, U_{\{\{i,j\}\}}$ mutually independent $Unif[0,1]$ random variables.

For binary undirected graphs without self-loops, the representation simplifies to:

$$A_{ij} = A_{ji} = \mathbb{1}(U_{\{i,j\}} \leqslant W(U_i, U_j))$$

for random W.

SBM's RDPG's etc all can be written in this form.

None-uniqueness of representation

For a model of the form:

$$A_{ij} = A_{ji} = \mathbb{1}(\eta_{ij} \leqslant w(\xi_i, \xi_j))$$

$(\eta_{ij})_{1\leqslant i<j\leqslant n} \sim Uniform[0,1]$, $\xi_1, ..., \xi_n \sim Uniform[0,1]$

We have that if we consider $\varphi(x)$ such that $\varphi(u) \sim Uniform[0,1]$ (measure-perserving transformation) then

$$(A_{i,j})_{1\leqslant i<j\leqslant n} \stackrel{d}{=} (\mathbb{1}(\eta_{ij} \leqslant w(\varphi(\xi_i), \varphi(\xi_j))))_{1\leqslant i,j\leqslant n}$$

Graphons also arise when one consider limits of dense graph sequences.

What do we mean by "limits" of graphs?

- Convergence of subgraph frequencies

- Convergence of some norm

- Convergence of subgraph probabilities, where vertices are sampled from a larger graph

- Convergence of cuts

- Convergence of spectra, etc

It turns out that, when properly defined, the first three are equivalent.

It turns out limit object is a graphon.

Subgraph frequencies

It turns out that there are several equivalent characterizations of subgraph frequency.

For fixed F (typically $V(F) \leqslant V(E)$), consider

$$t(F,G) = \frac{|hom(F,G)|}{|V(G)|^{|V(F)|}}$$

where $|hom(F,G)|$ is number of graph homomorphisms from F to G.

Equivalently,

$$t(F,G) = P(F \subseteq G[k])$$

where randomness comes from sampling $v_1, ..., v_k$ vertices with replacement, $G[k]$ is graph induced by $v_1, ..., v_k$.

Def Graph homomorphism

A graph homomorphism $f : V(F) \mapsto V(G)$ is a function that preserve adjacency.

MATH 586 Fall 2023
Statistics for Networks

Lecture 28: Graphon: Convergence of Subgraph Frequencies

Lecturer: Robert Lunde Scribe: Adrian Cao

28.60 Review

28.60.1 Exchangeability

The array $(A_{ij})_{1 \leq i,j, \leq n}$ is **jointly exchangeable** if for any permutation $\sigma : [n] \mapsto [n]$,

$$(A_{\sigma(i),\sigma(j)}) \stackrel{d}{=} (X_1, ..., X_n)$$

The (infinite) array $(A_{i,j})_{i,j \in \mathbf{N}}$ is **jointly exchangeable** if for all bijections $\sigma : \mathbf{N} \mapsto \mathbf{N}$

$$(A_{\sigma(i),\sigma(j)})_{i,j \in \mathbf{N}} \stackrel{d}{=} (A_{i,j})_{i,j \in \mathbf{N}}$$

28.60.2 Aldous-Hoover Theorem

Any jointly exchangeable infinite array admits representation

$$(A_{ij})_{i,j \in \mathbf{N}} \stackrel{d}{=} (f(U, U_i, U_j, U_{\{\{i,j\}\}})$$

For binary infinite arrays corresponding to undirected graphs without self-loops, the representation simplifies to:

$$(A_{ij})_{i,j \in \mathbf{N}} \stackrel{d}{=} (\mathbf{1}(U_{\{i,j\}} \leqslant \omega(U_i, U_j)))_{i,j \in \mathbf{N}}$$

for random ω.

28.60.3 Subgraph Frequency

Let

$$t(F,G) = \frac{|\hom(F,G)|}{|V(G)|^{|V(F)|}}$$

where $|\hom(F,G)|$ is number of graph homomorphisms from $V(F)$ to $V(G)$. (Graph homomorphisms preserve adjacency)

Equivalent to

$$t(F,G) = P(F \subseteq G[k])$$

where $k = |v(F)|$, $v_1, ..., v_k$ drawn with replacement from $1, \ldots, n$.

28.61 More Notation for Subgraph Frequency

28.61.1 Injected Homomorphisms and Induced Homomorphisms

One can also consider

$$t_{\text{inj}}(F,G) = P(F \subseteq G[k]')$$

where $v_1, ..., v_k$ drawn without replacement.

It turns out $|t(F,G) - t_{\text{inj}}(F,G)| \leq \frac{v^2(F)}{2v(G)}$, so t and t_{inj} share similar information.

Also, we have

$$t_{\text{ind}}(F,G) = P(F = G[k]'),$$

we have the relations $t_{\text{inj}}(F,G) = \sum_{|V(F)'|=V(F),F' \supseteq F} t_{\text{ind}}(F,G)$.

Recall the inclusion-exclusion principle

$$|\bigcup_{i=1}^{n} A_i| = \sum_{\emptyset \neq J, J \subseteq 1,\ldots,n} (-1)^{|J|+1} |\cap_{j \in J} A_j|,$$

then we have

$$t_{\text{ind}}(F,G) = \sum_{F' \supseteq F, V(F')=V(F)} (-1)^{e(F')-e(F)} t_{\text{inj}}(F,G).$$

Mainly, it says t_{inj} and t_{inj} shares the same amount of information.

28.61.2 Convergence

One possible notion of convergence is thus

$$\lim_{n \to \infty} t(F, G_n) = t(F, \omega) \quad \forall \text{ simple graphs } F$$

If it happens, it turns out limiting object is a graphon.

For intuition, suppose we map $A \in \{0,1\}^{n\times n}$ into a function $\omega_n : [0,1]^2 \to [0,1]$. To define ω_n, divide $[0,1]$ into intervals $\underbrace{[0,\frac{1}{n}]}_{I_1}, \underbrace{[\frac{1}{n},\frac{2}{n}]}_{I_2}, \ldots, \underbrace{[\frac{n-1}{n},1]}_{I_n}$. I_1 correspond

$$\omega_n(x_i,x_j) \underbrace{\prod_{i\in V(F)} dx_i}_{\text{permutation of nodes}} = \frac{|\hom(F,G)|}{n^k}, \quad n = |V(G)|, k = |V(F)|$$

If we take an analysis approach to limits, this suggests studying norms on functions of the form: $f : [0,1]^2 \to [0,1]$ds to node 1, I_2 corresponds to node 2, etc.

For $x \in [\frac{k-1}{n}, \frac{k}{n}], y \in [\frac{j-1}{n}, \frac{j}{n}]$, set $\omega_n(x,y) = A_{kj}$.

It turns out that

$$t(F,G) = \int \prod_{(i,j)}$$

MATH 586	Fall 2023
Statistics for Networks	
Lecture 29: Graphon	
Lecturer: Robert Lunde	Scribe: Zongxi Yu

29.62 Review

29.62.1 Graph limits

Consider a sequence of graphs $(t_n)_{n\geq 1}$, $n \to \infty$, under what conditions does the graph sequence "converge"?

One notion of convergence, convergence of homomorphism densities .

Suppose $\lim_{n\to\infty} t(F,G_n) = t(F,\omega) \quad \forall$ simple graphs F.

Then it will turns out that there exists a graphon $w : [0,1]^2 \to [0,1]$ such that $t(F) = t(F,\omega)$.

$t(F,G) = \frac{|\hom(F,G)|}{|V(G)|^{|V(F)|}}$, "homomorphism density"

Equivalent to $P(F \subseteq G[k])$ where $k = |v(F)|$, $v_1, ..., v_k$ are samples with replacement from $1, \ldots, n$.

Also equivalent $t(F,G) = \int \prod_{(i,j)\in E(F)} \omega_n(x_i,x_j) \prod_{i\in V(F)} dx_i$ where $\omega_n(x_i,x_j)$ is empirical graphon.

29.63 Cut Norm

If turns out the "right" norm to consider is the cut norm.

For a $n \times n$ matrix A, $||A||_\square = \frac{1}{n^2} \max_{s,t\in[0,n]} |\sum_{i\in S, j\in T} A_{ij}|$

It is clear that:

$||A||_\square \leq ||A||_1 \leq ||A||_2 \leq ||A||_\infty$

$||A||_1 = \frac{1}{n}\sum_{i,j}|A_{ij}|$

$||A||_2 = \sqrt{\frac{1}{n^2}\sum_{i,j}|A_{ij}^2|}$

$||A||_\infty = \max_{i,j}|A_{i,j}|$

Cut norm for functions:

$||\omega||_\square = \sup_{S,T\in[0,1]}|\int_{S\times T}\omega(x,y)d\,x d\,y|$

Also equivalent to:

$||\omega||_\square = \sup_{||f||_\infty\leq 1,||g||_\infty\leq 1}\|\int_{S\times T}\omega(x,y)f(x)g(y)d\,x d\,y\|$

We also have $||\omega||_\square \leq ||\omega||_1 \leq ||\omega||_2 \leq |\omega||_\infty$

To construct a metric, using cut norm, we want to align two kernels.

Consider $\delta_\square(\mu,\omega) = \inf_{\phi\in s[0,1]} d_\square(\mu,\omega^\phi)$ where $d_\square(u,v) = ||u-v||_\square$ and $w^\phi = w(\phi(x),\phi(y))$

29.63.1 Relationship between cut metric and subgraph frequencies

Theorem For any single graph F,
a) $|t(F,t_1) - t(F,t_2)| \leq d_\square(t_1,t_2)$
b) if $|t(F,t_1) - t(F,t_2)| \leq 4|E(F)|\delta_\square(t_1,t_2)$, then $\delta(t1,t2) \leq \frac{22}{\sqrt{log_2 k}}$

Theorem(function version) Let ω,$\omega' \in W$, where W is a class of function mapping $[0,1]^2$ to $[0,1]$,$c = \max(1,||\omega||_\infty,||w'||_\infty)$.
a) $t(F,\omega) - t(F,\omega') \leq 4mC^{m-1}d_\square(w,w')$, $m = |E(F)|$.
b) If $|t(F,\omega) - t(F,\omega')| \leq 3^{-12}$ for every F on k nodes, $d_\square(w,w') \leq \frac{22c}{log_c k}$

29.64 Szemeredi Regularity lemma

Function version idea:
Graphon's can be approximated by SBM's.

Define stepping operator, $w_p(x,y) = \frac{1}{\lambda(s_i)\lambda(s_j)}\int_{s_i\times s_j}\omega(x,y)dxdy$, where s_1,...s_k are a partition of $[0,1]$.

For every $\omega \in \omega_1$ and $k \leq 1$, $\exists$exists a partition p into at most k sets with positive measure for which $||\omega - \omega_p||_\square \leq \frac{2}{\sqrt{log k}}$.

Consider $\hat{\omega}$ space of functions: ω:$[0,1]^2 \to [0,1]$ equivalence classes for functions with $d_\square(\omega,\omega') = 0$

Theorem $(\hat{\omega}, d_\square)$ is compact.

MATH 586	Fall 2023
Statistics for Networks	
Lecture 30	
Lecturer: Robert Lunde	Scribe: Tianwei Zhou

30.65 Review

- For any simple F, $t(F, G_n = \mathbb{P}\big(F \subseteq G[k]\big)$, where $v_1, \ldots, v_n$ sample with replacement. One natural notion of convergence for a sequence of graphs $(G_n)_{n\geq 1}$:

$$\lim_n t(F, G_n) = t(F), \ \forall \text{ simple } F. \tag{30.8}$$

  It turns out that $t(F) = t(F, W)$ for some graphon W.

- Cut Norm for functions $w : [0,1]^2 \to [0,1]$.

$$\|W\|_\square = \sup_{S,T\subseteq[0,1]} \int_{S\times T} W(x,y)dxdy. \tag{30.9}$$

  And we have $\|W\|_\square \leq \|W\|_1 \leq \|W\|_2 \leq \|W\|_\infty$.

- Relationships between subgraph frequences and cut norm. Define distance

$$\delta_D(U, W) = \inf_{\varphi\in S[0,1]} \|U - W^\varphi\|_\square, \tag{30.10}$$

  where $W^\varphi = W\Big(\varphi(x), \varphi(y)\Big)$ and $S[0,1]$ denotes space of measure preserving transformation.

  Lemma 30.65.1 (Counting Lemma)**.** *For any single graph F, let W, W' mapping $[0,1]^2 \to [0,1]$.*

$$|t(F, W) - t(F, W')| \leq e(F)\delta_\square(W, W'). \tag{30.11}$$

  Lemma 30.65.2 (Inverse Counting Lemma)**.** *Let k be positive integer, U, W mapping $[0,1]^2 \to [0,1]$. Assume for every simple graph on k nodes $|t(F, W) - t(F, U)| \leq 2^{-k^2}$. Then*

$$\delta_D(U, W) \leq \frac{50}{\sqrt{\log k}}. \tag{30.12}$$

30.66 Theory of Graph Limit

For a partitions of vertices $\mathcal{P} = \bigcup_{i=1}^k S_i$, define stepping operator

$$W_\mathcal{P}(X, Y) = \frac{1}{\lambda(S_i)\lambda(S_j)} \int_{S_i\times S_j} W(x,y)dxdy, \tag{30.13}$$

where $X \in S_i$ and $Y \in S_j$.

Lemma 30.66.1 (Weak Regularity Lemma)**.** *For any $W \in W_1$ ($W : [0,1]^2 \to [-1,1]$), there exists $\mathcal{P}$ into at most k sets with positive measure for which*

$$\|W - W_\mathcal{P}\|_\square \leq \epsilon \tag{30.14}$$

Lemma 30.66.2. *For any $\epsilon > 0$, graphon W, partition $\mathcal{P}_0$, there exists a refinement $\mathcal{P}$ of $\mathcal{P}_0$ into no more than $4^{1/\epsilon^2}$ parts such that*

$$\|W - W_\mathcal{P}\|_\square \leq \epsilon. \tag{30.15}$$

Theorem 30.66.3. $\left(\tilde{W}_0, \delta_D\right)$ *is compact.*

Proof. It suffices to show for any sequence $W_n : [0,1]^2 \to [0,1]$, there exists a subsequence converging to an element in $\tilde{W}_0$.

For every n, k construct partitions $\mathcal{P}_{n,k}$ with corresponding stepping operator $W_{n,k}$ such that

1. $\|W_n - W_{n,k}\|_\square \leq \frac{1}{k}$.
2. $\mathcal{P}_{n,k+1}$ refines $\mathcal{P}_{n,k}$.
3. $|\mathcal{P}_{n,k}| = m_k$ depends only on k.

Idea: for each n, rearrange partitions with measure-preserving transformations so that partitions are intervals. For each k, pick subsequences such that

1. Each of the lengths of intervals are convergent.
2. Stepping operator defined on converging partition is converging.

Thus, for each k, we can pick subsequence n_j such that

$$\|W_{n_j,k} - U_k\|_\square \to 0, \tag{30.16}$$

where U_k is a limit for refinement level k. We thus have limit objects $U_1, U_2, \ldots$. Furthermore, one can verify that U_{k+1} is a refinement of U_k.
Now, pick $x, y \sim Uniform[0,1]$ and consider $U_1(x,y)$, $U_2(x,y)$.

Proposition 30.66.4. $\left(U_k(x,y)\right)_{k\geq 1}$ *is a martingale.*

Def: X_n is a martingale if

$$\mathbb{E}(X_n | X_{n-1}, \ldots, X_0) = X_{n-1}.$$

Martingale convergence theorem: Suppose $(X_n)_{n\geq 0}$ is a bounded martingale sequence. Then $X_n \xrightarrow{a.s.} X$ for some limitting X.
Thus by martingale convergence theorem and Proposition 30.66.4, $U_k(x,y) \xrightarrow{a.s.} U$.

Now, by triangle inequality, the proof can be finished by picking a subsequence such that

$$\delta_D(W_{n_j}, U) \leq \delta_D(W_{n_j}, W_{n_j,k}) + \delta_D(W_{n_j,k}, U_k) + \delta_D(U_k, U). \tag{30.17}$$

□

MATH 586 Fall 2023
Statistics for Networks

Lecture 31: Graph Limit

Lecturer: Robert Lunde Scribe: Tong Li

31.67 Review

- Suppose we have a sequence of graphs $(G_n)_{n\geq 1}$, we say graph converges when

$$\lim_n t(F, G_n) = t(F),\ \forall \text{ simple } F \tag{31.18}$$

,where $t(F, G_n) = \mathbb{P}\big(F \subseteq G[k]\big)$, where $v_1, \ldots, v_n$ sample with replacement.

- The "right" norm to consider is Cut Norm.

$$\|W\|_\square = \sup_{S,T\subseteq[0,1]} |\int_{S\times T} W(x,y)d\,x d\,y|. \tag{31.19}$$

And we have $\|W\|_\square \leq \|W\|_1 \leq \|W\|_2 \leq \|W\|_\infty$.
Define cut distance, let

$$\delta_\square(U, W) = \inf_{\varphi\in S[0,1]} \|U - W^\varphi\|_\square \tag{31.20}$$

where $W^\varphi = W\Big(\varphi(x), \varphi(y)\Big)$ and $S[0,1]$ denotes space of measure preserving transformation.

- We say $U \sim W$ if $\delta_\square(U, W) = 0$;
Consider space $\tilde{W}_0$: the quotient space of $W : [0,1]^2 \to [0,1]$, where two graphs are equivalent if $\delta_\square(U, W) = 0$. ($\tilde{W}_0$ be obtained from W_0 by identifying graphons with cut distance zero).
We have the compactness of the space $\big(\tilde{W}_0, \delta_D\big)$:

$\big(\tilde{W}_0, \delta_D\big)$ is compact.

31.68 Deviations in Homomorphism Densities

Corollary (31.68.1). For any $(G_n)_{n\geq 1}$, s.t. $t(F, G_n) \to t(F)$. For any simple F, there exists a graphon W s.t. $t(F) = t(F, W)$.

Proof. Since $\big(\tilde{W}_0, \delta_D\big)$ is compact, we can have convergent subsequences with limit W. By counting lemma,

$$|t(F, G_n) - t(F, W)| \leq e(F)\delta_\square(G_n, W).$$

Since $t(F, G_n)$ is Cauchy, $t(F, G_n) \to t(F)$. □

Definition 31.68.1. W-random Graph: Let $G_n(W)$ be the n-node graph generated by $A_{ji} = A_{ij} \sim Bernoulli\big(W(\xi_i, \xi_j)\big)$, where $\xi_1, \ldots, \xi_n \overset{\text{i.i.d}}{\sim} \mathbb{U}[0,1]$.

Theorem 31.68.2 (Exponentially Small Deviations in Homomorphism Densities)**.** *For any simple graph F,*

$$P(|t(F, G_n(W)) - t(F, W)| > \epsilon) \leq 2\exp\frac{-\epsilon^2 n}{4k^2} \tag{31.21}$$

which is saying that $t(F, G_n(W)) \xrightarrow{p} t(F, W)$

$$\frac{|hom(F, Gn(W))|}{n^{|V(F)|}} = \frac{1}{n^k}\sum_{V_1,\ldots,V_k} \mathbb{1}(F \subseteq G_n(k)) \tag{31.22}$$

Proof can start from the bounded-difference inequality which is the exponential-deviation inequality for functions of independent random variables, then connect the expected and limiting homomorphism density to get the desired inequality.

Proposition 31.68.3. *For each F,*

$$t(F, G_n(W)) \xrightarrow{a.s.} t(F, W) \tag{31.23}$$

Proof. We introduce Borel-Cantelli Lemma at first: Let $(E_n)_{n\geq 1}$ be sequence of events with the sum of the probabilities is finite, that is $\sum_{n=1}^{\infty} P(E_n) < \infty$, then the probability that infinitely many of them occur is 0: $P(E_n, io) = 0$, where io denotes infinitely often. We now consider the probabilities in the theorem are summable, by Borel-Cantelli Lemma, $|t(F, G_n(W)) - t(F, W)| > \epsilon$ is finitely often with probability 1. Then we have the convergence almost surely. □

Proposition 31.68.4. *For all homomorphism densities, Prop.1 tells us $t(F, G_n(W))$ almost surely convergents to $t(F, W)$ (that is $\forall F, P(t(F, G_n(W)) \to t(f, W)) = 1$), now we claim:*

$$G_n(W) \xrightarrow{a.s.} W \tag{31.24}$$

In other words,

$$P(\forall F, (t(F, G_n(W)) \to t(f, W))) = 1 \tag{31.25}$$

Proof. Consider the complementary event. Denote B as the event where not all of the homomorphism densities coverage and B_f is the failure-to-converge event for motif f. Then $P(B) \leq P(\cup_f B_f) \leq \sum_f P(B_f)$. By Proposition 1, $t(F, G_n(W)) \xrightarrow{a.s.} t(F, W)$. So $P(B_f) = 0$ for each f. Thus, $0 \leq P(B) \leq 0$, that is $P(B) = 0$. Therefore, $P(\forall F, (t(F, G_n(W)) \to t(f, W))) = 1$ [1]. □

31.69 Discussion

Problem with graph limit framework:

- Theory not interesting for sparse graph sequences (Consider sparse Erdős–Rényi model, $t(F, G_n) \to 0$)
- Functions of the form $W : [0, 1]^2 \to [0, 1]$ not rich enough to model heavy-tailed degree distribution.

MATH 586 Fall 2023
Statistics for Networks

Lecture 32: Sparse Graph Limits

Lecturer: Robert Lunde Scribe: Tong Li

32.70 Review

- Suppose we have a sequence of graphs $(G_n)_{n\geq 1}$, we say graph converges when

$$\lim_n t(F, G_n) = t(F), \ \forall \text{ simple } F \tag{32.26}$$

,where $t(F, G_n) = \mathbb{P}\big(F \subseteq G[k]\big)$, where $v_1, \ldots, v_n$ sample with replacement.
Counting Lemma: For any single graph F, let W, W' mapping $[0,1]^2 \to [0,1]$.

$$|t(F, W) - t(F, W')| \leq e(F)\delta_{\square}(W, W'). \tag{32.27}$$

Cut Distance:

$$\delta_{\square}(U, W) = \inf_{\varphi \in S[0,1]} \|U - W^{\varphi}\|_{\square} \tag{32.28}$$

where $W^{\varphi} = W\Big(\varphi(x), \varphi(y)\Big)$ and $S[0,1]$ denotes space of measure preserving transformation. $U \sim W$ if $\delta_{\square}(U, W) = 0$;

- Consider space $\tilde{W}_0$: the quotient space of $W : [0,1]^2 \to [0,1]$, where two graphs are equivalent if $\delta_{\square}(U, W) = 0$. ($\tilde{W}_0$ be obtained from W_0 by identifying graphons with cut distance zero).
  We have the compactness of the space $\big(\tilde{W}_0, \delta_D\big)$:

  $\big(\tilde{W}_0, \delta_D\big)$ is compact.

- For any $(G_n)_{n \geq 1}$, s.t. $t(F, G_n) \to t(F)$. For any simple F, there exists a graphon W s.t. $t(F) = t(F, W)$.

32.71 Approaches to Sparse Graph Limits

Continuing from the discussion we had at the last class, let's consider an example:
For sparse graph sequences where $P(G[2]) \to 0$, we have that $\lim_{n\to\infty} t(F, G_n) = t(F, 0), \forall \text{simple} F$.
So the graph limits in the last class are only for dense graph sequences; there are some modifications we need to discuss for the sparse graph case.

Corollary (32.71.1) ((sparse) L^p graphons). Idea: embed graphs in L^p instead of L^{∞}, normalize by edge density.
For sparse graph/function sequences, consider the distance

$$\delta_{\square}(W, W') = \delta_{\square}(\frac{W}{\|W\|_1}, \frac{W'}{\|W'\|_1}) \tag{32.29}$$

Comments: the intuition to do this is for a simple graph G, an upper bound on $\frac{G}{\|G\|_1} = \|G\|_1^{\frac{1}{p}-1}$ corresponds to a lower bound on $\|G\|_1$, which force G to be dense.

Theorem 32.71.1 (limits for L^p upper regular sequences)**.** *Let $(G_n)_{n\geq 1}$ be a sequence of graphs that L^p upper regular, then $\exists$ a subsequence G_{nj}, graphon W satisfying $\|W\|_p \leq C$, s.t.*

$$\delta_{\square}(\frac{G_{nj}}{\|G_{nj}\|_1}, W) \to 0 \tag{32.30}$$

Theorem 32.71.2. *If $\rho > 0$, satisfies $\rho_n \to 0$ and $n\rho_n \to \infty$ as $n \to \infty$, then*

$$\delta_{\square}(\rho_n^{-1} G_n(W, \rho_n), W) \xrightarrow{a.s.} 0 \tag{32.31}$$

32.72 Counting Lemma for L^p graphons

In this section, we discuss the extension of subgraph counts on the sparse case. In dense graph limits, homomorphism densities characterize convergence under the defined cut metric, while this left convergence cannot be maintained in sparse graph limits.

Theorem 32.72.1. *Let F be a simple subgraph with m vertices, max degree Δ. Let $\Delta < p < \infty$, If U and W are graphons with $\|U\|_p \leq 1$, $\|W\|_p \leq 1$, and $\delta_\square(U,W) \leq \varepsilon$, then*

$$|t(F,U) - t(F,W)| \leq 2m(m-1+p-\Delta)\big(\frac{2\varepsilon}{p-\Delta}\big)^{\frac{p-\Delta}{p-\Delta+m-1}} \tag{32.32}$$

32.73 Exchangeability

Another theory of Sparse Graph Sequences is related to graphexes. They consider exchangeability for adjacency measure.

$$\xi = \sum_{(X,Y)\in e(G)} \delta(X,Y) \tag{32.33}$$

An example of Kallenberg exchangeable graph[1].

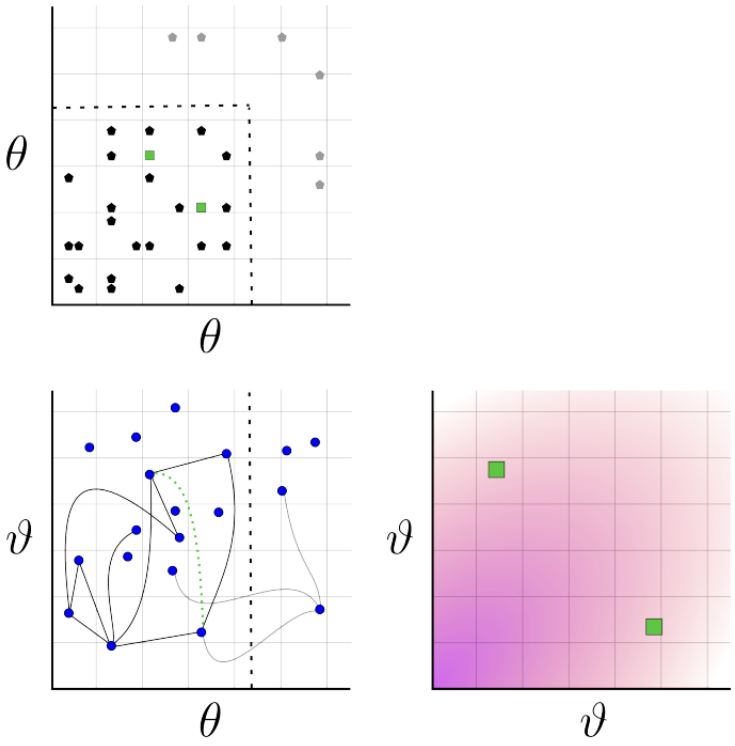


Figure 32.14: Kallenberg exchangeable graph

Theorem 32.73.1. *Let $\varphi : [0,\infty) \to [0,\infty)$ be a measure-preserving transformation, consider the process Y^{φ} where*

$$(\varphi(X), \varphi(Y)) \in Y^{\varphi} \Leftrightarrow (X, Y) \in Y \tag{32.34}$$

Y is exchangeable if $Y^{\varphi} \stackrel{d}{=} Y$.

A representation theorem for exchangeable point process suggests model:

$$Z_i j|(\theta_k, \gamma_k)_{k=1,2,,\ldots} = \text{Bernoulli}(W(\gamma_i, \gamma_j)) \tag{32.35}$$

where $W : [0,\infty)^2 \to [0,1]$
(θ_k, γ_k) is unit rate Poisson process.

Bibliography

[1] Kolaczyk, Eric D. *Statistical Analysis of Network Data Methods and Models*, Springer (Springer Series in Statistics), 2009.

[2] Norris, James. *Markov Chains*, Cambridge University Press, 1997.

[3] Veitch, Victor, and Daniel M. Roy. "The class of random graphs arising from exchangeable random measures." arXiv preprint arXiv:1512.03099 (2015).