Ronak Mehta
Dr. Joshua Vogelstein
TA: Eric Bridgeford

Neuro Data Design
Analysis of Network Data
Reading Notes

Chapter 1

Main Ideas

This short chapter motivates the study of networks with examples, and explains the format of the book.

(a) Networks are represented with differing levels of formality, from sets of interconnected "things" to formal mathematical graphs.

(b) Immediate issues associated with this kind of data: relational nature, high-dimensionality, and dependence.

(c) Questions we might ask: How big is the network? What does it look like? What is the traffic going through it?

(d) In network analysis, descriptive statistics (network mapping) is much more non-trivial task than in traditional statistics.

Important Metrics, Algorithms, or Results

None in this introductory chapter.

Implications

The chapter briefly mentions biological networks, which is interest to us (connectomes). This can generalize to other biomolecular processes.

Chapter 2

Main Ideas

This chapter introduces the graph theoretic background of network science.

Basic Graph Theory

(a) The chapter introduces definitions of a graph $G = (V, E)$, subgraph, induced subgraph, order $N_V = |V|$, and size $N_E = |E|$.

(b) We assume for now simple graphs without loops or multi-edges. These would be multi-graphs.

(c) Several notions of connectivity: adjacency of edges or vertices, incidence of a vertex on an edge, vertex degree.

(d) Pre-known definitions: reachable, path, walk, connected, digraph, diameter, tree, forest, ancestor, descendant, leaf.

(e) Degree Sequence: sequence formed by arranging vertex degrees in non-decreasing order.

(f) Strongly connected: digraph has all points reachable from one another; weakly connected: underlying graph is connected.

(g) Geodesic distance: number of edges in shortest path between vertices. This generalizes with edge weights, being a special case of $w_e = 1$.

(h) Connected graph: graph where every vertex is connected to every other vertex. A clique is a connected subgraph.

(i) Regular graph: graph with every vertex the same degree. Common to model lattice structures.

(j) Directed Acyclic Graph (DAG) is not the same as a directed tree.

(k) Bipartite Graph: $G = (V, E)$ is bipartite if we can split V into disjoint sets V_1 and V_2 such that every edge has one endpoint in V_1 and the other in V_2. We can use these for matching networks, or membership networks.

(l) Planar graph: graphs that can be embedded in a plane without edges crossing. Used for networks with spatial component.

Matrix Algebra

The operations on simple adjacency matrix A show properties of the graph, but with an easier representation.

(a) Row sum i = degree of node i. For a directed adj. matrix, the $d_{i,\text{out}}$ is the row sum i.

(b) If we exponentiate A^r, then $(A^r)_{ij}$ is the number of walks from vertex i to j of length r.

(c) The n by m incidence matrix B has $B_{ij} = 1$ if vertex i is incident on edge j.

(d) D is the diagonal matrix containing the degree sequence. We have $BB^T = D - A$.

(e) $L = D - A$ is the Laplacian of a graph. We have $x^T Lx = \sum_{(i,j)\in E} (x_i - x_j)^2$, show the distance of points mapped to by functions on graphs.

(f) The eigenvectors and eigenvalues of L reveal structural properties of the graph. A large value of the second to smallest eigenvalue represents a largely "connected" graph.

Graph Data Structures

(a) A graph can be represented on computers as an adjacency matrix, adjacency list, or incidence list.

(b) While a matrix is easier to work with, it is more computationally expensive. Sparse graphs have less than half of their adjacency matrices filled.

(c) BFS, DFS, algorithmic complexity.

Probability and Statistics

The following section only contains topics that I haven't seen in depth yet.

(a) Markov Chains: integer-valued random variables called "state", that can go between each other with a transition probability.

(b) This eventually converges to a stationary distribution, which can be given by the 1-eigenvector of the transition matrix.

(c) Given sequence of events $A_1, A_2, ..., A_n$. A_n occurs with high probability of $\lim_{n\to\infty} P(A_n) = 1$.

(d) The Bonferonni Correction divides the desired α by m for m tests. Unfortunately, this gives the test low power.

(e) From a Bayesian approach, the ridge estimator is defined as $^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y$, with $\lambda = \frac{\sigma^2}{\tau^2}$. τ^2 is the prior variance, so this will become the MLE estimator if the prior variance goes to infinity.

(f) Model selection procedures commonly fall into the class of complexity penalized least squares.

Connection to Graphs

(a) We can measure the gene expression levels $x \in \mathbb{R}^d$ of N_V genes, and take their pairwise correlations. When the correlation is high enough, we can all them "connected" and draw an edge between those genes. Here we must learn the threshold $\hat{\rho}_{ij} > t$ to determine if an edge exists. This is a hypothesis testing setup.

(b) We can regress on the feature vectors X_i, X_j to determine whether edge Y_{ij} exists, as well as the other edges in the network. These are called auto-regression models.

(c) Lastly, a traffic flow matrix can be used to predict the traffic on a set of edges given a current state of edge traffic.

Implications

While a lot of the probability and statistics information is the same, models need to take into account the relational nature of the data. Learning a function to predict the existence of an edge seems to be a very useful approach that we can use on a brain graph, to identify connectivity patterns. Similarly, the Markov Chain info might be useful to model the EEG data, and do comparative analysis on EEG scans, similar to the ARMA ideas explored last semester.

Chapter 3

Main Ideas

Important Metrics, Algorithms, or Results

Implications

Chapter 4

Main Ideas

Important Metrics, Algorithms, or Results

Implications

Chapter 5

Main Ideas

Important Metrics, Algorithms, or Results

Implications

Chapter 6

Main Ideas

Important Metrics, Algorithms, or Results

Implications

Chapter 7

Main Ideas

Important Metrics, Algorithms, or Results

Implications

Chapter 8

Main Ideas

Important Metrics, Algorithms, or Results

Implications

Chapter 9

Main Ideas

Important Metrics, Algorithms, or Results

Implications

Chapter 10

Main Ideas

Important Metrics, Algorithms, or Results

Implications