Math 168: Network Science Discussion Notes

Yacoub Kureh

Spring 2018

These notes are meant to accompany Professor Mason Porter's lectures and my discussions. Please let me know if you find any mistakes or have questions.

1 Introduction

Being from the social media generation, we probably agree —at least on some level—that networks are important to study. However, by the end of this course, we'll start seeing that networks are everywhere and crucial to our understanding of the world around us.

In this course, we are going to use many ideas from mathematics and other sciences, but we'll try to make it as self-contained as possible. We'll start by (re)viewing some of the rudiments of graph theory which sets the stage for network science.

1.1 Terminology from Graph Theory

A *graph* G is an ordered pair (V, E) where V is the vertex or node set and E is the edge set which consists of 2-element subsets of V.

TO DO: Insert image of a basic graph.

To be more precise, the above definition is for an *undirected, simple graph*. It is *undirected* in the sense that the edges do not have a 'from–to' direction (E consists of sets, not ordered pairs) and it is *simple* meaning there is precisely zero or one edge between distinct vertices and no edge linking a vertex to itself (E is itself a set consisting of 2-element subsets). It's usually understood that when one says graph they mean undirected and simple unless otherwise stated or made clear by context. These simple objects are incredibly interesting to study in their own right. Let's point out some language real quick: distinct vertices connected by an edge are called 'adjacent' and distinct edges that share a vertex are called 'incident.'

Let's consider the graph G_{Stu} where the vertices are students in this class and there is an edge between two students if they have taken at least one class together before this class (if we included this class then there would be an edge between all vertices, which isn't as interesting). We might ask, who is 'popular' in this graph? One way to answer this question might use degrees. The *degree* of a vertex v (denoted $d(v)$) is the number of edges incident to v. So, one might say the vertex or vertices with the greatest degree are the most popular as they have shared the most courses with other students. Let's put the question of popularity to the side for now, and focus on what else we can say about degrees.

What if we know the degrees of all the vertices in our graph? How well do we know the graph? There are two equivalent ways to store this information about degrees. We can arrange all the degrees as a list in non-increasing order which we'll call the *degree sequence*. Or we can consider the *degree distribution* of a graph, which is the function

$$p_{\text{deg}}(k) = \frac{|\{v \in V : d(v) = k\}|}{|V|}.$$

Let's examine the degree distribution of some real-world graphs.

Insert: Examples of degree distributions.

We can learn a lot about a graph from its degrees, but not everything. It is important to recognize that one cannot recover complete information about a graph from just its degree sequence.[1] As an aside, a very interesting 'inverse problem' is asking what sequences can be degree sequences for graphs. In 1960, Erdos and Gallai proved that the following conditions are both sufficient and necessary for a list of non-negative integers $d_1 \geq d_2 \geq \ldots \geq d_n$ to be a degree sequence: $d_1 + d_2 + \ldots + d_n$ is even and

$$\sum_{i=1}^{k} d_i \leq k(k-1) + \sum_{i=k+1}^{n} \min(d_i, k)$$

holds for every k in $\{1, 2, \ldots, n\}$. You should check the meaning of the first condition for yourself as an exercise, hint: try to relate the sum of degrees to $|E|$.

Consider our G_{Stu} graph again, another question we might ask is if it is 'connected'? Our intuitive real-world definition coincides here, but let's also give the precise definition. First, define a *walk* to be a alternating sequence of vertices and edges beginning and ending on vertices

$$v_{i_1}, e_{j_1}, v_{i_2}, e_{j_2}, \ldots, e_{j_n}, v_{i_{n+1}}$$

such that each edge is incident to the vertex before and after it. The number of edges involved in a walk is its length, so this was a walk of length n. A *path* is a walk which does not repeat edges or vertices except for the possibility that $v_{i_1} = v_{i_{n+1}}$ in which case it is a *cycle*.[2] We call a graph *connected* if there is a path between every pair of vertices. Not all graphs are

---

[1] To make this more precise, we'll need to define isomorphisms on graphs, but in short, there are non-isomorphic graphs with the same degree sequence.

[2] Unfortunately, not everyone agrees on these definitions, so *caveat lector*.

connected of course, and how to study such graphs can be a challenging question.

It may be feasible to draw a pictorial representation of G_{Stu} and attempt to answer most questions we have by hand, however, we'll soon want the aid of a computer. Defining a graph as an ordered pair of sets is useful formally, but from a computational perspective, it is often helpful to store data about a graph in an *adjacency matrix*, which is a square binary (0 or 1 entries) matrix of size $|V|$ where the entry A_{ij} represents whether an edge between vertices i and j exists or not. In simple graphs, $A_{ii} = 0$ for all i. There are many uses for the adjacency matrix.

1.1.1 Exercises

- What term do we have for the quantity $(A^2)_{ii}$?
- Check that $(A^k)_{ij}$ counts the number of walks of length k from vertex i to vertex j, where $(A^k)_{ij}$ denotes the (i, j)th entry of the kth power of the matrix A, i.e. A times itself k times.
- Convince yourself that if a graph is connected, for each i and j there should be a k that makes $(A^k)_{ij}$ positive, but note that it may not stay positive as you increase k.

1.2 Networks

Now we have a definition for a graph, but what is a network? A *network* is an object which consists of vertices representing entities which are connected by edges representing ties between them. This is a less mathematically formal definition than the one we gave for graph but it highlights the essence. By this definition, a graph is an example of a network. In fact, directed graphs (where edges are ordered pairs rather than sets) are networks, pseudographs (or non-simple graphs in which loops and multiple edges are permitted) are networks, hypergraphs (where we allow hyperedges connecting two or more vertices) are networks, and weighted graphs (where edges have an assigned weight) are networks, etc.. We'll discuss these and more in greater detail later on as we explore more real networks, but for now you can probably begin to imagine how these additional structures will help us model the world around us.

Suppose we are given a network to study; we are well prepared to start asking many interesting problems about networks. What is, if there is any, the community structure of the network? Does the network exhibit clustering? Does the network exhibit the small-world property? Is this network similar to another network? We might even try to make inferences or predictions from the network.

Before concluding this introduction, it would be remiss not to mention dynamics. So far, the networks mentioned have been fairly static, but real-world networks evolve. A hugely important field in network science is the interplay between dynamics and networks. Typically we say there are two kinds of dynamics: dynamics on networks and dynamics of networks. We can also couple these two kinds of dynamics and see how the interact.

2 Background

2.1 Statistics

It's quite likely that for your projects, you'll be working with data. Luckily for you there is a lot of data that you have access to. Sometimes the data will be organized in way that makes it easily amenable to the tools of network science. Other times, it may require a lot of cleaning and cleverness. Regardless, it will be useful to have some of the basic tools from statistics to help analyze the data.

Given N samples $x_1, \ldots, x_N$, we define the k^{th} *sample moment* to be

$$\frac{1}{N}\sum_{i=1}^{N} x_i^k.$$

The first sample moment is known as the *mean*, which will also be denoted as $\langle x \rangle$. We define the standard deviation as

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \langle x \rangle)^2}.$$

In addition to defining mean, moments, and standard deviation for samples, we can define them for random variables. For example the k^{th} *moment* for a discrete random variable X with support $\{x_i\}_{i \in I}$ is

$$\sum_{i \in I} x_i^k p_i$$

where $p_i = P(X = x_i)$. For a continuous random variable Y the k^{th} *moment* is

$$\int_{-\infty}^{\infty} y^k f(y) dy$$

where $f(y)$ is the probability density function.

Beyond these measures, there is the important question of how do we fit sample data to a distribution, i.e. is the data normally distributed? We won't be able to cover the topic of fitting here (both for reasons of time and it's quite a challenging topic in itself).

2.2 Scale-free Networks

Consider a network whose degree distribution fits to a pareto distribution, also known as power-law distribution. That is, the degree distribution is well approximated by power-law

$$p_{\text{deg}}(k) \sim k^{-\gamma}.$$

These networks are sometimes called scale-free because they do not exhibit a 'characteristic scale'. There is contention around this terminology, which I'll try to summarize and then give references for further reading.

Let's start with Barabasi. For Barabasi, "a network that has a power-law degree distribution, regardless of any other structure, is called a scale-free network.Ïn Chapter 4 of his *Network Science* book, he explains the terminology using moments. You've probably heard of the heuristic 68–95–99.7 rule that helps you remember what percent of normally distributed data lie within a sigma interval around the mean; e.g. approximately

95% of normally distributed data should be within two standard deviations of the mean. A similar result can be stated for the Poisson or other unimodal distributions. The takeaway is that if the degrees of a network are distributed this way, then there is some sense of the scale as most degrees are comparable to each other and so the mean is understood to be the scale. However, for a power-law distribution, the distribution mean exists only if $\gamma > 2$, and the distribution standard deviation only if $\gamma > 3$. That is to say, if $\gamma < 3$, then the standard deviation for the distribution is infinite (check the integral for the second moment diverges). In Barabasi's words, this "means that when we randomly, choose a node, we do not know what to expect: The selected node's degree could be tiny or arbitrarily large. Hence networks with $\gamma < 3$ do not have a meaningful internal scale, but are 'scale-free'."

However, "power-law and scale-free are very different conceptsäccording to Chung and Lu who discuss scale-free networks in Section 3.5 in their *Complex Graphs and Networks* book. For them, the discussion of scales involves a discussion of both space and time. There are scale-free in space networks which exhibit self-similarity which they go on to define and there are scale-free in time networks which they also go on to define, but I'll try to summarize what they say: Suppose you have a process for generating a network by adding nodes and edges one at a time. Next, suppose you can divide up time into equal-length intervals and combine all the nodes born in the same interval into a single super-node. If the resulting graph still has a power-law degree distribution with the same exponent regardless of the size of the intervals, it is said to be scale-free in time (note: it will have fewer nodes).

For more reading on this topic:

1. Mathematical results on scale-free random graphs by Bol-

lobas
`https://www.stat.berkeley.edu/~aldous/Networks/boll1.pdf`

2. Scale-free networks are rare by Broido and Clauset
`https://arxiv.org/pdf/1801.03400.pdf`

3. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications
`https://people.csail.mit.edu/jshun/6886-s18/papers/LADW06.pdf`

2.3 Graph Theory

Let's add a few more concepts to our repertoire. Every graph we'll consider here is undirected, unweighted, and simple. The *complete graph on* n *vertices*, denoted K_n is a graph where every distinct pair of vertices is connected by an edge. That is, all vertices are adjacent and the graph has $\frac{n(n-1)}{2}$ edges. A *bipartite* graph is one which the vertex set V can be decomposed into two disjoint sets V_1 and V_2 such that no pair of vertices from the same set are adjacent. The *complete bipartite graph of* p, q *vertices*, denoted $K_{p,q}$, is a bipartite graph where $|V_1| = p$ and $|V_2| = q$ and every vertex in V_1 is adjacent to every vertex in V_2. You should check that $K_{p,q}$ has pq edges.

Very often, we visualize our graphs by drawing dots and lines[1] on a plane, dots representing the vertices and lines representing the edges. A natural question we might ask ourselves, 'is it possible to draw every graph without the lines crossing?', to which the answer is no. A graph G is said to be *planar* if the

---

[1] *Dots and Lines* is also the title of a graph theory textbook by Richard J. Trudeau

graph can be embedded in (or, in simple English, drawn on) the plane such that edges do not cross. For example, K_1, K_2, K_3 and K_4 are planar. K_5 is not planar. The three utilities problem is a classic math puzzle where you have three houses and three utilities gas, electricity, and water, and your goal is draw a line from every house to every utility without crossing. In graphic theoretic terms, we are asked to give a planar embedding of $K_{3,3}$ without edges crossing, which is not possible[2]. These two graphs, K_5 and $K_{3,3}$, play a significant role in planarity testing thanks to two theorems. Kuratowski's theorem states that a finite graph is planar if and only if it does not contain a subgraph that is a subdivision of the complete graph K_5 or the complete bipartite graph $K_{3,3}$ (utility graph). Wagner's theorem states that a finite graph is planar if and only if it does not have K_5 or $K_{3,3}$ as a minor. Next, we consider the notion of distance. The *geodesic distance,* or simply the *distance* between two vertices in a graph is the length of a shortest path between them. For example, in K_n, the distance between any pair of distinct vertices is 1. In a friendship network, friends of friends (who themselves are not friends) are a distance 2 apart. The *diameter* of a graph is maximum distance in a graph; that is, it is the longest shortest path between any two vertices. The diameter gives us global information about the graph about distance. Another global is the *mean geodesic distance*, which is the mean over all distances taken over all pairs of distinct vertices. The diameter is already greater than or equal to the mean geodesic distance.

You may be familiar with the small world phenomenon of strangers being linked by a short chain of friends. We can now define a small-world network. A *small-world network* refers to an ensemble of networks in which the mean geodesic distance between nodes L grows at most proportionally to the logarithm of

---

[2] It is possible to do so on a mug though!

the number of nodes, i.e. $L = O(log(N))$ where N is the number of nodes. Note, you'll often see the term used to describe a single network in this family. Note that Newman defines the *small-world effect* in the same way but replaces the mean geodesic distance with the graph diameter which is more restrictive.

There are many ways to store information about a graph in a matrix. We've already seen the adjacency matrix and we now introduce the Laplacian matrix, which has many beautiful spectral properties. Unfortunately, there isn't consensus on how to define the Laplacian, but we'll go with $L = D - A$, where D is the matrix with the vertex degrees on the diagonal and zeroes elsewhere and A is the adjacency matrix. That is,

$$L_{i,j} = \begin{cases} deg(i) & i = j \\ -1 & i \neq j, (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

[TODO Insert example]. You should check that the sum of across any row or any column in L is zero. What this implies is that the vector of all ones in $\mathbb{R}^N$ (where $|V| = N$) is an eigenvector of L with eigenvalue 0. L is notably symmetric and therefore can be diagonalized in an orthonormal basis and has real eigenvalues. There are physical interpretations to L that we may revisit later on, but for now we'll just discuss one connection between the dimension of the nullspace of L and the number of connected components of G (they're equal!). A *connected component* (or just component) of a graph is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the whole graph. [TODO Insert picture]

If a graph has c connected components, then with appropriate reordering its adjacency matrix consists is block diagonal and consequently so is the Laplacian. Each block in the Laplacian is in fact the Laplacian matrix for subgraph. We can now make c linearly independent vectors that live in the nullspace of L: for each component, there is a Laplacian block L_i in L, and for each block we consider the vector of all ones in the entries corresponding to the block and zeros elsewhere. Check that these vectors are indeed in the nullspace and linearly independent. Thus we have that c is at most the nullity of L. In the other direction, let's consider the vectors in $\mathbb{R}^N$ as real-valued weights assigned to each vertex. The quantity $x^T Lx$ will be of interest.

$$\begin{aligned} x^T Lx &= x^T(D-A)x \\ &= x^T Dx - x^T Ax \\ &= \sum_{v_i \in V} deg(v_i)x_i^2 - \sum_{(i,j)\in E} 2x_i x_j \\ &= \sum_{(i,j)\in E} x_i^2 + x_j^2 - \sum_{(i,j)\in E} 2x_i x_j \\ &= \sum_{(i,j)\in E} (x_i - x_j)^2 \end{aligned}$$

As an aside, this tells that L is positive semi-definite, so all of its eigenvalues are non-negative. Suppose L has nullity k, that is, L has a nullspace $N \subset \mathbb{R}^N$ of dimension k. For all $x \in N$, $x^T Lx = 0$, but this means $\sum_{(i,j)\in E}(x_i - x_j)^2 = 0$ and so $x_i = x_j$ if $(i,j) \in E$ and moreover $x_i = x_k$ if v_i and v_k are in the same component, and so x is constant on components. Thus the dimension of N could be at most the number of components. This completes the proof.

3 Random Graphs

4 Clustering Coefficient

5 Homework Questions

In this chapter, I'll provide partial solutions to past homework assignments based on requests.

5.1 Problem 8.2 from Newman's *Networks*

We are given that the degree distribution is $p_k = Ce^{\lambda k}$ where C and λ are constants. For (a), we want to find C as a function of λ. What this question is asking: what does C need to be in order for p_k to be a probability function. It's not explicitly stated, but we'll take the support to be from $k = 0$ to infinity. You could consider other supports though. The condition that will give us C is that $\sum_{i=0}^{\infty} p_i = 1$. Solving this gives us that $C = 1 - e^{-\lambda}$. For (b), we want the fraction P of vertices that have degree k or greater. We know that p_k tells us the fraction of vertices that have exactly degree k, so

$$P = \sum_{i=k}^{\infty} p_i = \sum_{i=k}^{\infty} Ce^{-\lambda i} = C \sum_{i=k}^{\infty} (e^{-\lambda})^i.$$

This geometric sum comes out to be $e^{-\lambda k}$.
For (c), we want the fraction W of ends of edges that are attached to vertices of degree k or greater. The reference given in the textbook is to a paper which discuss the same kind of calculation in the context of wealth inequality. You'll sometimes

hear the richest 20% of individuals have 80% of the money.[1] The analogy here is that the 'richest' (in terms of high degree) x% of the vertices 'own' (i.e. are incident to) y% of the edges where $y > x$. So let's actually compute out the numbers. What we want in the numerator of our fraction are the total number of edge ends attached to vertices of degree k or greater, and the denominator will be total number of edges ends. The denominator is thus twice the number of edges, but we don't have that so we'll just have to take whatever expression we get for the numerator and set $k = 0$. Let n denote the number of nodes, then np_k nodes have degree k and so knp_k is the total number of edge ends attached to nodes of degree k. Now we try to compute

$$\sum_{i=k}^{\infty} inp_i = n\sum_{i=k}^{\infty} ip_i = n\sum_{i=k}^{\infty} i(1-e^{-\lambda})e^{-\lambda i} = n(1-e^{-\lambda})\sum_{i=k}^{\infty} ie^{-\lambda i}.$$

To compute $\sum_{i=k}^{\infty} ie^{-\lambda i}$, we can notice it is equal to the negative of the derivative of a geometric sum, i.e. $\sum_{i=k}^{\infty} ie^{-\lambda i} = -\frac{d}{d\lambda}\sum_{i=k}^{\infty} e^{-\lambda i}$. We then use the result from (b) and have that

$$-\frac{d}{d\lambda}\sum_{i=k}^{\infty} e^{-\lambda i} = -\frac{d}{d\lambda}\frac{e^{-\lambda k}}{1-e^{-\lambda}} = -\frac{e^{-\lambda k}(-k)(1-e^{-\lambda}) - e^{-\lambda k}e^{-\lambda}}{(1-e^{-\lambda})^2}.$$

So substituting and simplifying, we get that

$$\sum_{i=k}^{\infty} inp_i = ne^{-\lambda k}\frac{k(1-e^{-\lambda}) + e^{-\lambda}}{1-e^{-\lambda}}.$$

---

[1] It's apparently way worse than that: http://fortune.com/2017/11/14/credit-suisse-millionaires-millennials-inequality/

Let's plug 0 in for k to get the total number of edge ends, $n\frac{e^{-\lambda}}{1-e^{-\lambda}} = \frac{n}{e^{\lambda}-1}$. Now W is just the ratio, so this gives

$$W = \frac{ne^{-\lambda k}\frac{k(1-e^{-\lambda})+e^{-\lambda}}{1-e^{-\lambda}}}{\frac{n}{e^{\lambda}-1}} = e^{-\lambda k}[k(e^{\lambda}-1)+1].$$

For (d), you're asked to show the Lorenz curve is given by $W = P + \frac{1-e^{\lambda}}{\lambda} P \ln P$. This follows from substitution and algebra. Note, the Lorenz curve just means writing W in terms of P, the form depends on the distribution, and will look different for different distributions.

6 References and Acknowledgements

Thank you to Mason Porter for sharing his slide decks and notes. *Networks: An Introduction* by Mark Newman is the primary textbook for this course. *Network Science* by Albert-Laszlo Barabasi is another useful resource and is available for free here: `http://networksciencebook.com`. *Complex Graphs and Networks* by Fan Chung and Linyuan Lu is another excellent resource.