Math 168 – Intro to Networks

University of California, Los Angeles

Duc Vu

Spring 2022

This is math 168 – Introduction to Networks taught by Professor Chodrow. We meet weekly on MWF from 10:00 am to 10:50 am for lecture. The required textbook for the class is *Networks* 2^{nd} by *Newman*. Other course notes can be found at my blog site. Please let me know through my email if you spot any typos in the note.

Contents

1 Lec 1: Mar 28, 2022 **3**
1.1 Introduction . . . 3

2 Lec 2: Mar 30, 2022 **4**
2.1 Networks and Matrices . . . 4

3 Lec 3: Apr 1, 2022 **6**
3.1 Measures and Metrics . . . 6

4 Lec 4: Apr 5, 2022 **8**
4.1 Measures and Metrics (Cont'd) . . . 8

5 Lec 5: Apr 6, 2022 **10**
5.1 Erdos-Renyi Random Graph . . . 10

6 Lec 6: Apr 8, 2022 **12**
6.1 Paths and Branching Processes in ER Random Graphs . . . 12

7 Lec 7: Apr 11, 2022 **13**
7.1 Giant Component in Sparse Erdos-Renyi . . . 13

8 Lec 8: Apr 13, 2022 **14**
8.1 Experimental Lecture on ER Theory . . . 14

9 Lec 9: Apr 15, 2022 **15**
9.1 Configuration Model . . . 15

10 Dis 1: Mar 29, 2022 **17**
10.1 Review of Linear Algebra . . . 17
10.2 Review of Probability . . . 18
10.3 Some Useful Inequalities . . . 20

List of Theorems

4.7 Laplacian Formula for Cuts 9
10.7 Spectral for Real Matrices 18
10.8 Perron-Frobenius Version 1 18
10.9 Perron-Frobenius Version 2 18

List of Definitions

2.1 Graph 4
2.2 Adjacency Matrix 4
2.4 Walk 4
3.1 Degree 6
3.2 Path-connected 6
3.3 Geodesic Path 6
3.4 Diameter 7
4.1 Triadic Closure 8
4.2 Local Clustering Coefficient 8
4.4 Laplacian Matrix 8
4.5 Clustering/Partition 9
4.6 Cut Value 9
5.1 Random Graph 10
5.4 Cycle 11
6.1 Path 12
6.2 Branching Process 12
7.1 Giant Component 13
9.1 Degree Sequence 15
9.2 Configuration Model Random Graph 15
10.2 Matrix Kernel 17
10.3 Range 17
10.4 Eigenvalue 17
10.5 Spectral Radius 17
10.6 Span 17
10.10 Joint Probability 18
10.11 Conditional Probability 18
10.13 Independent Events 19
10.14 Expectation 19

§1 Lec 1: Mar 28, 2022

§1.1 Introduction

Some introduction and logistics stuffs of the class. Nothing mathy is discussed in this lecture.

§2 Lec 2: Mar 30, 2022

§2.1 Networks and Matrices

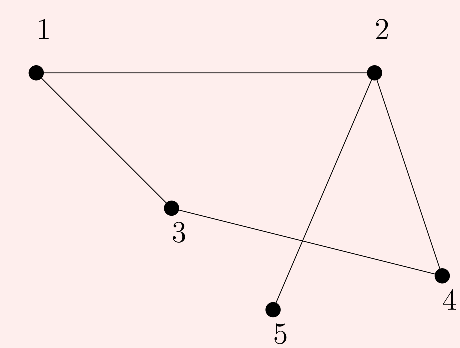
Definition 2.1 (Graph) — A (simple, undirected) graph is $G = (N, E)$, a node set N and an edge set $E \subseteq N \times N$ s.t. $i \neq j \forall (i, j) \in E$.

Definition 2.2 (Adjacency Matrix) — The adjacency matrix $\mathbf{A}$ of a graph $G = (N, E)$ is a matrix in $\mathbb{R}^{n\times n}$ where $n = |N|$ with entries

$$a_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

Example 2.3

Consider the following graph



The adjacency matrix is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

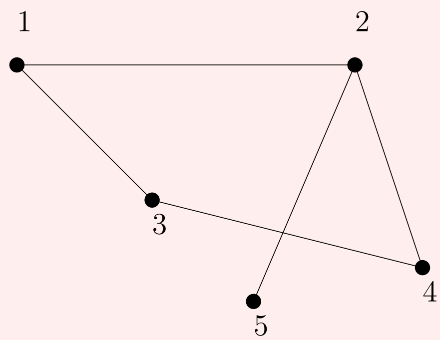
Definition 2.4 (Walk) — A walk in $G = (N, E)$ is a sequence of edges $(i_1, j_1), (i_2, j_2), \ldots, (i_k, j_k)$ where

$$j_1 = i_2, j_2 = i_3, \ldots, j_{k-1} = i_k$$

This is a walk of length k (number of edges) from i_1 to j_k.

Example 2.5

Consider the above example



Walk $3 \to 2$ of length

- $1 : \emptyset$
- $2 : (3,1), (1,2); (3,4), (4,2)$
- $3 : \emptyset$
- $4 : (3,1), (1,2), (2,1), (1,2)$

Fact 2.1. The $ijth$ entry of $\mathbf{A}$ counts the number of walks of length 1 from node i to j.

Conjecture 2.1. The $ijth$ entry of $\mathbf{A}^k$ counts the number of walks of length k from i to j.

Proof. Suppose inductively that $W(k) \triangleq \mathbf{A}^k$ has entries $w_{ij}(k)$ counting k-walks from $i \to j$. Consider $W(k+1) = W(k)A$. Its entries are

$$w_{ij}(k+1) = \sum_{l \in N} w_{il}(k) a_{lj}$$

$\square$

§3 Lec 3: Apr 1, 2022

§3.1 Measures and Metrics

A walk of length 2, $i \leftrightarrow i$, is the number of edges attached to node $i \triangleq$ degree of node i, k_i

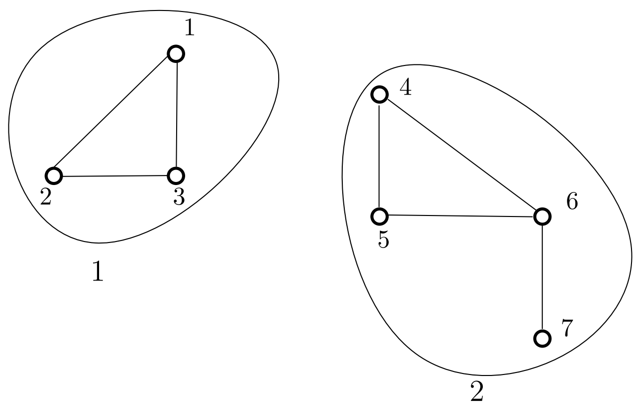
$$k_i = \sum_{j\in N} a_{ij} = \sum_{j\in N} a_{ji} = ii\text{th entry of } \mathbf{A}^2$$

Definition 3.1 (Degree) — The degree k_i of a node i is the number of edges attached to it

$$k_i = |\{j : (i,j) \in E\}|$$

Definition 3.2 (Path-connected) — Nodes i and j are path-connected if $\exists$ a walk $i \leftrightarrow j$ of any length. The connected component of i is the set of notes to which i is path-connected.
G is connected if it has 1 connected component.

Consider a disconnected graph G



Then, the adjacency graph is

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & 0 \\ 0 & \mathbf{A}_2 \end{bmatrix} \text{ up to permutations of node labels}$$

Question 3.1. How big is a graph?

- Number of nodes
- Number of edges
- Diameter

Definition 3.3 (Geodesic Path) — Geodesic (shortest) path between $i \leftrightarrow j$ is a walk s.t. no walk has shorter length.

Definition 3.4 (Diameter) — Diameter of G is

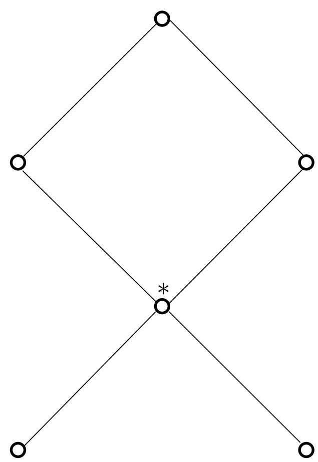
$$\max_{ij} \text{geodesic distance}(i,j)$$

which is undefined if i and j are not connected.

Example 3.5

"6 degrees of separation": in social networks, the diameter is usually about 6.

Node Importance:

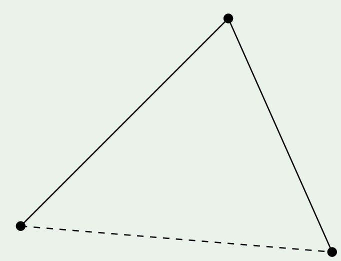


- * has highest degree
- If * were removed, graph would be disconnected
- Short average distance to other nodes
- Betweeness: # of geodesic paths passing through $i \in N$.

§4 Lec 4: Apr 5, 2022

§4.1 Measures and Metrics (Cont'd)

Definition 4.1 (Triadic Closure) — In networks, the observation that this phenomenon



happens a lot is called triadic closure.

Calculation: # of triangles attached to node i = # of walks of length 3 $i \leftrightarrow i$ (A^3 diagonal) $\cdot \frac{1}{2} = \frac{1}{2} \sum_j \sum_k a_{ij} a_{jk} a_{ki}$.
To compute the # of possible triangles attached to i

1. Calculate k_i
2. $\binom{k_i}{2}$

   Exercise 4.1. Express in terms of the adjacency matrix $\mathbf{A}$.

Definition 4.2 (Local Clustering Coefficient) — The local clustering coefficient CC_i at node i is

$$\frac{\text{\# of triangles at } i}{\text{\# of possible triangles}}$$

Note: $0 \le CC_i \le 1$.

Remark 4.3. On average, CC_i is high (many triangles can be observed) and global measures are high.

Definition 4.4 (Laplacian Matrix) — The (combinatorial) Laplacian matrix of a graph $\mathbf{L} \in \mathbb{R}^{n\times n}$

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

where

$$\mathbf{D} = \begin{bmatrix} k_1 & & & \\ & k_2 & & \\ & & \ddots & \\ & & & k_n \end{bmatrix}$$

Definition 4.5 (Clustering/Partition) — A clustering/partition of a graph is a partition of N, $\{C_1, C_2, \dots, C_l\}$

$$N = \bigcup_{j=1}^{l} C_j, \qquad C_j \cap C_{j'} = \emptyset \text{ if } j \neq j'$$

Let $c_i \triangleq$ cluster of node i.

Definition 4.6 (Cut Value) — The cut value of a partition $\{C_1, \dots, C_l\}$ is

$$\frac{1}{2} \sum_{i,j \in N} a_{ij} \underbrace{\mathbb{1}[c_i \neq c_j]}_{\substack{=1,\ c_i \neq c_j \\ =0 \text{ otherwise}}}$$

Idea: Good clustering have small cut values.
Setting: 2 clusters $\{C_1, C_2\}$

$$s \in \mathbb{R}^n, \quad s_i = \begin{cases} +1, & c_i = 1 \\ -1, & c_i = 2 \end{cases}$$

Theorem 4.7 (Laplacian Formula for Cuts)
The cut value of $\{C_1, C_2\}$ is $\frac{1}{4} s^\top \mathbf{L} s$.

Proof. Consider

$$\begin{aligned} x^\top \mathbf{L} x &= x^\top (\mathbf{D} - \mathbf{A}) x \\ &= \sum_{i \in N} k_i x_i^2 - \sum_{i,j \in N} a_{ij} x_i x_j \\ &= \sum_{i,j \in N} a_{ij} x_i^2 - \sum_{i,j \in N} a_{ij} x_i x_j \\ &= \frac{1}{2} \left(\sum_{i,j \in N} a_{ij} x_i^2 + \sum_{i,j \in N} a_{ij} x_j^2 - 2 \sum_{i,j \in N} a_{ij} x_i x_j \right) \\ &= \frac{1}{2} \sum_{i,j \in N} a_{ij} (x_i - x_j)^2 \end{aligned}$$

So

$$\begin{aligned} s^\top \mathbf{L} s &= \frac{1}{2} \sum_{i,j \in N} a_{ij} (s_i - s_j)^2 \\ &= \frac{1}{2} \sum_{i,j \in N} a_{ij} \left(4 \mathbb{1}[c_i \neq c_j] \right) \end{aligned}$$

□

§5 Lec 5: Apr 6, 2022

§5.1 Erdos-Renyi Random Graph

Definition 5.1 (Random Graph) — Random graph is a probability distribution over graphs.

Definition 5.2 — An Erdos-Renyi random graph on n nodes with edges probability p is written $G(n, p)$.

To sample, we take each pair of nodes and draw an edge between them i.i.d with probability p.

Question 5.1. How many pairs are there?

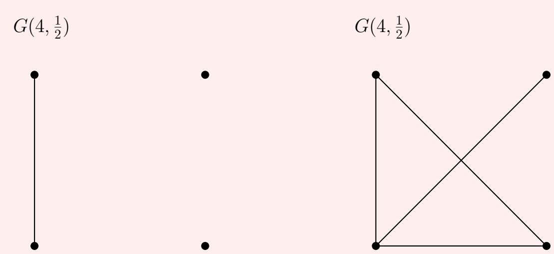
There are $\binom{n}{2}$. Also,

$$\mathbb{E}\left[\# \text{ of edges}\right] = p\binom{n}{2}$$

So # edges $\sim$ Binomial$\left(\binom{n}{2}, p\right)$.

Example 5.3

Consider



Question 5.2. What is the average degree c?

We can see that $c = \frac{1}{2}$ for the left graph and $c = 2$ for the right graph.

Note: The average degree can be calculate as $c = \frac{2m}{n}$ where m is the number of edges and n is the number of nodes.

Expected # of Triangles in ER

$$\mathbb{E}\left[\# \text{ of triangles}\right] = \binom{n}{3}p^3$$

where each edge is independent. Recall the global clustering coefficient is

$$C = \frac{\# \text{ of triangles} \cdot 3}{\# \text{ of wedges}}$$

where a wedge is a graph with 3 nodes and 2 edges. Then,

$$\mathbb{E}\left[\# \text{ of wedges}\right] = 3\binom{n}{3}p^2$$

Note that

$$\mathbb{E}\left[C\right] \neq \frac{\mathbb{E}\left[\triangle\right] \cdot 3}{\mathbb{E}\left[\#\text{ of wedges}\right]} = p$$

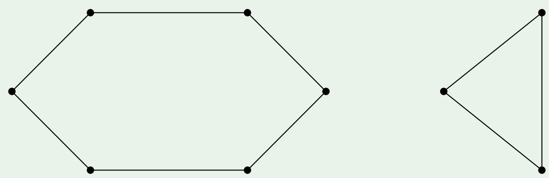
Fact 5.1. As $n \to \infty$, $C \overset{\text{in dist.}}{\longrightarrow} p$.

Node degree in ER

$$\mathbb{E}\left[k_i\right] = (n-1)p = c$$

As p increase, $\mathbb{E}\left[k_i\right]$ also increases (linearly). However, this is not always the case. Consider sparse ER with $p = \frac{c}{n-1}$. So, as $n \to \infty$, $p = \frac{c}{n-1} \to 0$, i.e., sparse ER model has very little clustering.

Definition 5.4 (Cycle) — A cycle on node i is a walk from $i \leftrightarrow i$ with no repeated nodes or edges.



We can observe that 3-cycles are rare in large sparse ER ($n \to \infty$).

§6 Lec 6: Apr 8, 2022

§6.1 Paths and Branching Processes in ER Random Graphs

Definition 6.1 (Path) — A path is a walk with no node repetitions.

Path lengths: Pick $i, j \in N$. Define $R_k \triangleq \#$ of paths of length k between i and j. Let's compute $r_k = \mathbb{E}[R_k]$.

$$\mathbb{E}[R_k] \approx \mathbb{E}[\#(\text{ path to } l \text{ and } (l,j) \in E)] = \mathbb{E}[\# \text{ of length to } l \text{ of length } k-1]\, p = r_{k-1}p$$

So

$$r_k \approx \underbrace{r_{k-1}p}_{\text{for path through } l} \quad \underbrace{(n-2)}_{\#\text{ of ways to choose } l \neq i,j}$$

Also, notice that

$$r_k \approx r_{k-1}p(n-1) = r_{k-1}c = c^{k-1}r_1 = c^{k-1}p$$

Question 6.1. What length k makes path likely?

We have

$$\log r_k \approx (k-1)\log c + \underbrace{\log p}_{\log c - \log n}$$

$$k \approx \frac{\log r_k + \log n}{\log c}$$

Assume $r_k = 1$. Then, consider the world population of 8 billions with average degree of 1000

$$k \approx \frac{\log n}{\log c} \approx \frac{\log 8 \cdot 10^9}{\log 10^3} \approx 3.4$$

Notice that if $c \leq 1$, the expression above doesn't make any sense.
Galton-Waston Branching Process

Definition 6.2 (Branching Process) — Let p be a probability distribution on $\mathbb{Z}$, called the offspring distribution. A branching process with distribution p is a sequence of random variables $X_0, X_1, X_2, \ldots$ s.t. $X_0 = 1$ and for $t \geq 1$,

$$X_t = \sum_{i=1}^{X_{t-1}} Y_i$$

where each Y_i is distributed i.i.d. according to p.

Branching processes create tree-graphs without cycles, which we can utilize to better understand the behavior of ER random graph.

§7 Lec 7: Apr 11, 2022

§7.1 Giant Component in Sparse Erdos-Renyi

Fact 7.1. Say we have a Poisson(c) process, then

$$\mathbb{E}[X_k] = c^k$$

Total number of individuals in $\mathbb{E}$

$$\mathbb{E}\left[\sum_{k=0}^{\infty} X_k\right] = \sum_{k=0}^{\infty} \mathbb{E}[X_k] = \sum_{k=0}^{\infty} c^k = \begin{cases} \frac{1}{1-c} & 0 < c < 1 \\ \text{divergent “}\infty\text{”} & c \geq 1 \end{cases}$$

Let's consider

$$\begin{aligned} P\left(\frac{\text{size of component containing node } i}{n} > a\right) &= P(\text{size} > an) \\ &\leq \frac{\mathbb{E}[\text{size}]}{an} \qquad \text{(Markov's)} \\ &= \frac{1}{an}\frac{1}{1-c} \to 0 \text{ unless } a = 0 \end{aligned}$$

In the case of $a = 0$, $P\left(\frac{\text{size}}{n} > 0\right) = 1$.

Definition 7.1 (Giant Component) — A sequence $G\left(n, \frac{c}{n-1}\right)$ as $n \to \infty$ has a giant component(GC) if

$$P\left(\frac{\text{component containing random node } i}{n} > a(>0)\right) \geq b > 0$$

In other words,

$$\mathbb{E}[\text{size of largest component}] = an \text{ for some } 0 < a \leq 1$$

Fact 7.2. Sequence $G\left(n, \frac{c}{n-1}\right)$ has a giant component if and only if $c > 1$.

Let u be the probability P that a node is not in giant component, $s = 1 - u$ is probability that a node is in giant component, $sn =$ size of giant component.

$$u = \Bigg(\underbrace{1-p}_{\text{not connected}} + \underbrace{pu}_{\text{connected not in GC}}\Bigg)^{n-1}$$

Let's simplify the above expression.

$$\begin{aligned} u &= (1 - p(1-u))^{n-1} \\ &= \left(1 - \frac{c(1-u)}{n-1}\right)^{n-1} \\ &= e^{-c(1-u)} \text{ as } n \to \infty \end{aligned}$$

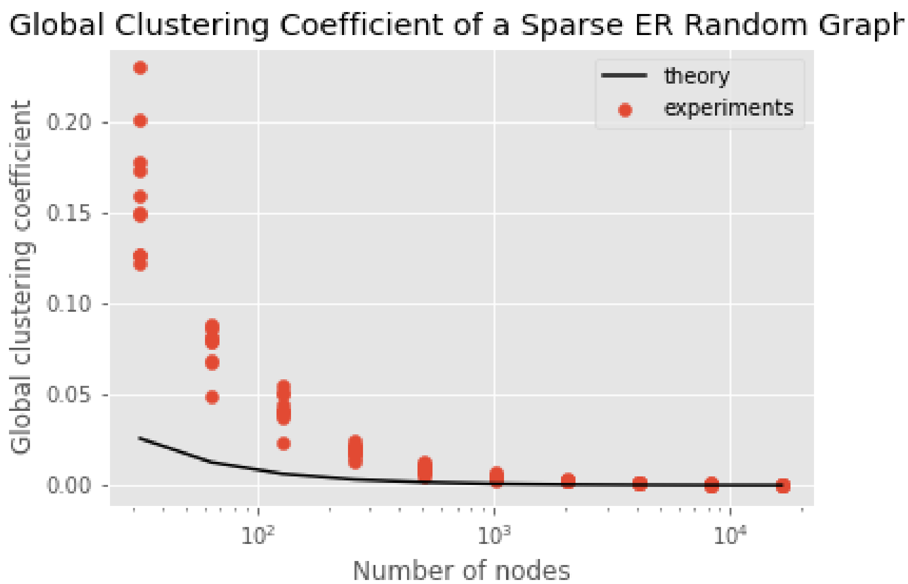
Replace $s = 1 - u$

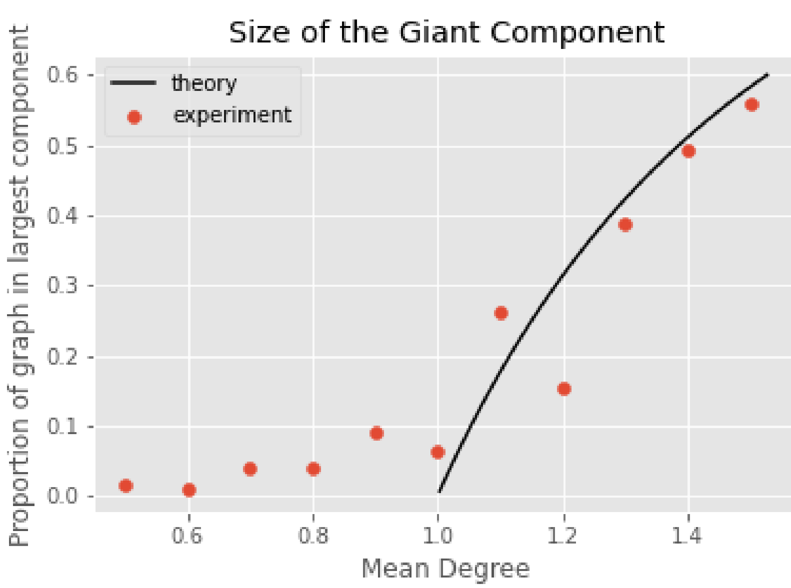
$$s = 1 - e^{-cs}$$

§8 Lec 8: Apr 13, 2022

§8.1 Experimental Lecture on ER Theory

In this lecture, we did some experiments with Python to check whether they agree with the theoretical results that we discussed previously on ER random graph.





§9 Lec 9: Apr 15, 2022

§9.1 Configuration Model

Definition 9.1 (Degree Sequence) — The degree sequence of $G = (N, E)$ with $|N| = n$ is $\vec{k} \in Z^n$ s.t. degree of $i \in N = k_i$.

Definition 9.2 (Configuration Model Random Graph) — The configuration model random graph with degree sequence $\vec{k}$ is a uniformly random graph among all graph with degree sequence $\vec{k}$.

Stub-Matching:

Select uniformly random pairs of half edges and turn them into edges until we run out of edge pair (we then have a graph). However, this method is not perfect as we can have a problem with self-loop or parallel edges, i.e., we only want simple graphs.

Fact 9.1. For $n \to \infty$, if the degree sequence doesn't grow in its entries (sparsity), then $P\left(\text{simple graph}\right) > \varepsilon > 0$.

Fact 9.2. Stub matching (conditioned on getting a simple graph) samples from configuration model.

Moment of the degree sequence:

Definition 9.3 — Degree distribution $p_k = P\left(\text{random node has degree } k\right) = \frac{\#\text{ nodes of degree } k}{n}$.

The l^{th} moment is defined as

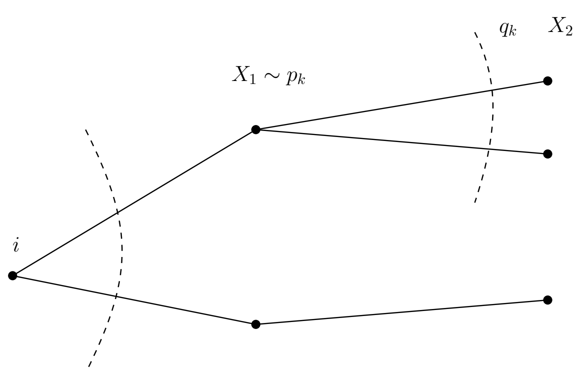
$$\langle k^l \rangle \triangleq \sum_k p_k k^l$$

So

$$\langle k^0 \rangle = \sum_k p_k = 1$$

$$\langle k^1 \rangle = \sum_k p_k k$$

Branching Process:



$$\mathbb{E}\left[X_1\right] = \langle k \rangle$$

and

$$\begin{aligned}
q_k &= P\left(\text{if I follow an edge to node j, the number of additional edges on j} = k\right)\\
q_k &= \frac{(k+1)p_{k+1}n}{\text{\# of half-edges=2m}}\\
&= \frac{(k+1)p_{k+1}}{\langle k\rangle}
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\text{\# offspring in 2nd gen from single parent}\right] &= \sum_{k=0} kq_k = \sum_{k=0}\frac{k(k+1)p_{k+1}}{\langle k\rangle}\\
&= \sum_{k'=1}\frac{(k'-1)k'p_{k'}}{\langle k\rangle}\\
&= \frac{1}{\langle k\rangle}\sum_{k'=1}\left(k'^2-k'\right)p_{k'}\\
&= \frac{1}{\langle k\rangle}\left(\langle k^2\rangle-\langle k\rangle\right)
\end{aligned}$$

Branching heuristic for giant component: Giant component iff $\frac{\langle k^2\rangle-\langle k\rangle}{\langle k\rangle} > 1$.

§10 Dis 1: Mar 29, 2022

§10.1 Review of Linear Algebra

Networks can be represented as ordinary matrices and the related graph Laplacian so linear algebra will be very important for us.

Example 10.1

Spectral graph theory techniques.

Definition 10.2 (Matrix Kernel) — The kernel of a matrix $\mathbf{A}$ denoted $\ker(\mathbf{A})$ is the set of vectors $\vec{x}$ s.t. $\mathbf{A}\vec{x} = \mathbf{0}$.

Definition 10.3 (Range) — The range of a matrix $\mathbf{A}$ denoted $\text{im}(\mathbf{A})$ is the set of vectors $\vec{x}$ s.t. there exists a vector $\vec{y}$ s.t. $\mathbf{A}\vec{y} = \vec{x}$.

Definition 10.4 (Eigenvalue) — λ is an eigenvalue of $n \times n$ matrix $\mathbf{A}$ with right-eigenvector (usually just called eigenvector if there's no chance of confusion) $\vec{x} \neq \mathbf{0}$ if $\mathbf{A}\vec{x} = \lambda\vec{x}$.
It has left-eigenvector $\vec{y} \neq \mathbf{0}$ if $\vec{y}^\top \mathbf{A} = \lambda \vec{y}^\top$ or equivalently $\mathbf{A}^\top \vec{y} = \lambda \vec{y}$.

Note: If $\vec{x}$ is a left or right eigenvector with eigenvalue λ, then so is $\alpha\vec{x}$ for any scalar $\alpha \neq 0$.

Definition 10.5 (Spectral Radius) — The spectral radius for an $n \times n$ matrix $\mathbf{A}$ denoted $\rho(\mathbf{A})$ is the maximum magnitude of its eigenvalues

$$\rho(\mathbf{A}) = \max\left\{|\lambda| : \lambda \text{ is an eigenvalue of } \mathbf{A}\right\}$$

Definition 10.6 (Span) — The span of a set of vectors is the set of all linear combinations of those vectors

$$\text{span}\left\{v_1, v_2, v_3\right\} = \left\{a_1 v_1 + a_2 v_2 + a_3 v_3\right\}$$

where a_1, a_2, a_3 are scalars.

Question 10.1. How do we calculate eigenvalues/eigenvectors?

1. Calculate the characteristics polynomial

$$p_{\mathbf{A}}(\lambda) = \det\left(\mathbf{A} - \lambda \mathbf{I}_n\right)$$

2. The eigenvalues are the roots $\lambda_1, \ldots, \lambda_n$ of the characteristic polynomial

3. Calculate the eigenvectors, i.e., we want to solve for $\vec{x}$ s.t. $(\mathbf{A} - \lambda\mathbf{I})\vec{x} = \mathbf{0}$.

Theorem 10.7 (Spectral for Real Matrices)
Let $\mathbf{A}$ be a $n \times n$ symmetric matrix. Then $\mathbf{A}$ has real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ (not necessarily distinct) and there exists corresponding eigenvectors $\{v_1, v_2, \ldots, v_n\}$ that form an orthonormal basis for $\mathbb{R}^n$.

Furthermore, $\mathbf{A}$ is diagonalizable here, i.e., $\mathbf{QDQ}^\top = \mathbf{A}$ where $\mathbf{D}$ is the diagonal matrix of eigenvalues and $\mathbf{Q}$ is its corresponding matrix of eigenvectors. This theorem also applies to the adjacency matrix of any undirected network since it will be symmetric.

Theorem 10.8 (Perron-Frobenius Version 1)
If $\mathbf{A}$ is a $n \times n$ real matrix with all non-negative entries, then $\mathbf{A}$ has a non-negative leading eigenvalue κ_1 with

$$\kappa_1 = \rho(\mathbf{A}) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } \mathbf{A}\}$$

So for this leading eigenvalue κ_1, all other eigenvalues λ of $\mathbf{A}$ satisfy $|\lambda| \le \kappa_1$.

In addition there exists left and right eigenvectors of κ_1 with all non-negative entries.

Proof. Hw 0 # 3a. □

Theorem 10.9 (Perron-Frobenius Version 2)
If $\mathbf{A}$ is the adjacency matrix (could be weighted) for a strongly-connected directed network (or any connected undirected network, then we call $\mathbf{A}$ irreducible). If $\mathbf{A}$ is a real $n \times n$ matrix that either

a) has all strictly positive entries or

b) is irreducible

then

i) The leading eigenvalue κ_1 with $\kappa_1 \ge |\lambda|$ is strictly positive ($\kappa_1 > 0$) and has one-dimensional eigenspace $\{v : \mathbf{A}v = \kappa_1 v\}$

ii) κ_1 has a leading left and right eigenvectors associated with it that have all strictly positive entries.

iii) The only non-negative eigenvectors of $\mathbf{A}$ are multiples of the leading eigenvectors. The eigenvectors for all other eigenvalues of $\mathbf{A}$ have at least one negative entry.

§10.2 Review of Probability

Definition 10.10 (Joint Probability) — The probability of events A and B occurring is denoted $P(A \cap B)$ or $P(A, B)$.

Definition 10.11 (Conditional Probability) — The probability of $\mathbf{A}$ given that $\mathbf{B}$ occurred is denoted $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Theorem 10.12
The probability of A or B (or both) occurring is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Law of Total Probability

Let X denote the sample space of all possible events. Let $\{B_i\}_{i=1}^{\infty}$ be a countable (it could be finite) partition of X, so $X = \bigcup_{i=1}^{\infty} B_i$, then

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i)$$

Bayes' Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Definition 10.13 (Independent Events) — Events A and B are independent if an only if one of the following equivalent statements hold

i) $P(A \cap B) = P(A)P(B)$

ii) $P(A|B) = P(A)$

iii) $P(B|A) = P(B)$

Definition 10.14 (Expectation) — We have two cases

- Discrete: Let X be a discrete random variable with possible events $\{x_i\}_{i=1}^{\infty}$ and probability mass function (PMF) p_x. The expected value of $g(X)$ is

$$E\left[g(X)\right] = \langle g(X) \rangle = \sum_{i=1}^{\infty} g(x_i)p_x(x_i)$$

- Continuous: Let X be a continuous random variable with probability density function (PDF) f_X. Then

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f_x(x)dx$$

Note: Expectation is linear, i.e.,

$$E\left[X+Y\right] = E\left[X\right] + E\left[Y\right]$$
$$E\left[\alpha X\right] = \alpha E\left[X\right]$$

Definition 10.15 — A random variable X has

1. Mean $\mu_x = E[X]$
2. Variance $\text{var}(X) = \sigma_x^2 = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E[X]^2$
3. n^{th} moment (about zero) $E[X^n]$

§10.3 Some Useful Inequalities

There are some useful inequalities to keep in mind

1. Cauchy-Schwartz for expectation
$$|E[XY]| \leq \sqrt{E[X^2]\,E[Y^2]}$$
2. Markov's Inequality: If X is a non-negative random variable and $a > 0$, then
$$P(X \geq a) \leq \frac{E[X]}{a}$$
3. Chebyshev's Inequality: Let X be a random variable with mean μ and variance σ^2. Then for $a > 0$,
$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$
4. Jensens's Inequality: If f is a convex function, then
$$E[f(X)] \geq f(E[X])$$
If f is a concave function, then the reverse inequality holds.

Note: $f : \Omega \to \mathbb{R}$ is convex if for all $0 \leq t \leq 1$ and $x_1, x_2 \in \Omega$,
$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$
f is concave if $-f$ is convex.