Lectures 2: Graph Theory and Social Networks

Alexander Wolitzky

MIT

6.207/14.15: Networks, Spring 2022

Plan

First part of the course focuses on the physical structure of networks, with no or very simple models of behavior.

Basic tool: **graph theory**, the mathematical study of graphs/networks.

- We use the terms "graph" and "network" interchangeably.

This lecture: Basic graph theory language and concepts for describing and measuring networks.

- Next week: more advanced concepts and applications. E.g., Google's PageRank algorithm, which ranks webpages by "importance" based on their position in the Web network.

Types of Networks in the Real World

A network is a set of units (**nodes** or **vertices**) connected by relationships (**links** or **edges**).

Types of networks:

- Social and economic networks: nodes are people or groups of people.
  - Friendship networks, business relationships between firms, intermarriages between families, employment relations in the labor market
- Information networks: nodes are "information objects"
  - Web links, citation network between academic articles, semantic/classification networks (e.g., taxonomies)
- ...

Types of Networks in the Real World (cntd.)

- Technological networks
  - Infrastructure networks like internet, power grid, transportation networks
  - Temporary networks like sensor networks, autonomous vehicles
- Biological networks
  - Food web, protein interaction network, neural network, network of metabolic pathways

History of Study of Graphs/Networks

Historical study of networks:

- Mathematical graph theory: central part of discrete math
  - Started with Euler's 1735 solution to the Königsberg bridge problem.
- Social network analysis in sociology.
  - Typical studies involved circulation of questionnaires, leading to relatively small networks; also little focus on individual behavior.

Recent years witnessed a substantial change in network research.

- From analysis of single small graphs ($<$100 nodes) to statistical properties of large-scale networks (millions/billions of nodes).
  - Motivated by availability of computers and computer data.
- On a different front, integration of game theory and graph/social network theory.[5]
  - Later in the course.

Graphs

An **graph** consists of a set of **nodes** $N = \{1, \ldots, n\}$ and an $n \times n$ matrix $g = [g_{ij}]_{i,j \in N}$ called the **adjacency matrix**, where $g_{ij} \in \{0, 1\}$ denotes the absence/presence of an edge from node i to node j.

- In a **weighted graph**, the edge weight $g_{ij} > 0$ can take on non-binary values, representing the intensity of the interaction.

In an **undirected graph**, $g_{ij} = g_{ji}$ for all $i, j \in N$ (g is symmetric).

- E.g. Facebook friends

In a **directed graph (digraph)**, g_{ij} and g_{ji} may differ.

- E.g. web links

Examples: draw the graphs corresponding to adjacency matrices:

- Example 1: $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ Example 2: $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

Graphs

Equivalently, can represent a graph by (N, E), where $E \subseteq N \times N$ is the set of edges.

- For directed graphs, E is the set of "directed" edges, write $(i, j) \in E$.
- For undirected graphs, E is the set of "undirected" edges, write $\{i, j\} \in E$.

Example 1: $E_d = \{(1,2), (2,3), (3,1)\}$
Example 2: could write as either $E_u = \{\{1,2\}, \{1,3\}, \{2,3\}\}$ or $E_d = \{(1,2), (2,1), (2,3), (3,2), (3,1), (1,3)\}$

We sometimes denote $g_{ij} = 1$ with the notation $(i, j) \in g$, or $\{i, j\} \in g$, or even $ij \in g$.

Walks, Paths, and Cycles

For an undirected graph (N, E):

- A **walk** is a sequence of edges $\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{K-1}, i_K\}$.
- A **path** between nodes i and j is a sequence of edges $\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{K-1}, i_K\}$ such that $i_1 = i$ and $i_K = j$, and each node in the sequence $i_1, \ldots, i_K$ is distinct. (i.e. a walk with no repeated nodes)
- A **cycle** is a path where the final node equals the initial node.
- A **geodesic** between nodes i and j is a "shortest path" (i.e. with minimum number of edges) between these nodes.

The **length** of a walk (or path) is the number of edges in the walk (or path).

- The **distance** between nodes i and j is the length of a geodesic between them (or ∞ if no such path exists).

For directed graphs, the same definitions hold with directed edges (in which case we say "a path **from** node i **to** node j").

Powers of the Adjacency Matrix

The powers of the adjacency matrix contain useful information about walks and paths.

Under the convention $g_{ii} = 0$, the matrix g^2 tells us the **number** of walks of length 2 between any two nodes:

$$\begin{aligned} (g \times g)_{ij} &= \sum_{k \in N} g_{ik} g_{kj} \\ &= \#\left\{k : \{i,k\}, \{k,j\} \text{ is a walk between } i \text{ and } j\right\} \end{aligned}$$

(since $g_{ik} g_{kj} = 1$ if $\{i,k\}, \{k,j\}$ is such a walk, $= 0$ otherwise).

Similarly, the matrix g^3 tells us the number of walks of length 3 between any two nodes:

$$\begin{aligned} \left(g^2 \times g\right)_{ij} &= \sum_{k_2 \in N} \left(g^2\right)_{ik_2} g_{k_2 j} \\ &= \#\left\{ \begin{array}{c} (k_1, k_2) : \{i,k_1\}, \{k_1,k_2\}, \{k_2,j\} \\ \text{is a walk between } i \text{ and } j \end{array} \right\}. \end{aligned}$$

Powers of the Adjacency Matrix (cntd.)

By induction, g^k tells us the number of walks of length k between any two nodes.

This also gives a useful way to express the distance between nodes i and j: it is the smallest integer k such that $\left(g^k\right)_{ij} \neq 0$.

A similar interpretation works for weighted graphs: given a weighted adjacency matrix g, $\left(g^k\right)_{ij}$ is the sum of the "values" of all length-k walks from i to j, where the value of a walk is the product of the weights on each link.

You'll see more ways of using the adjacency matrix on the pset.

Connectivity and Components

An undirected graph is **connected** if for every two nodes there exists a path between them.

A graph (N', E') is a **subgraph** of (N, E) if $N' \subset N$, $E' \subset E$, and $\{i, j\} \in E'$ implies $i, j \in N'$. (Each link must have ends.)

A **component** of a graph is a maximal connected subgraph.

- That is, a connected subgraph that is not contained in any larger connected subgraph.

An edge $\{i, j\}$ is a **bridge** if deleting it increases the number of components.

Note: the adjacency matrix of a graph with more than one component can be written in block-diagonal form: that is, the 1's are confined to square blocks along the diagonal, with all other elements equal to 0. (Convince yourself.)

Connectivity and Components in Directed Graphs

A directed graph is

- **connected** if the underlying undirected graph is connected (i.e. ignoring the directions of the edges).
- **strongly connected** if each node can reach every other node by a "directed path".

A **strongly connected component** is a maximal strongly connected subgraph. That is,

1. Each node in the subgraph can reach every other node in the subgraph by a directed path contained in the subgraph.
2. The subgraph is not contained in any larger subgraph with this property.

Directed Graphs (cntd.)

- The **out-component** of a set of nodes $S \subset N$ is the set of nodes $T \subset N$ that can be reached by a directed path starting from some node in S.
- The **in-component** of a set of nodes $S \subset N$ is the set of nodes $T \subset N$ that can reach some node in S by a directed path.

Note: the strongly connected component of a node i consists of the intersection of its out-component and its in-component.
Proof:

- Fix two nodes j and k in the intersection of i's out-component and in-component.
- j can reach i, because j is in i's in-component.
- i can reach k, because k is in i's out-component.
- So j can reach k (by a path through i).

Some Special Networks

- A **clique** (or **complete network**) is a graph where all nodes are linked to each other.
- A **tree** is a connected (undirected) graph with no cycles.
  - A connected graph is a tree if and only if it has $n-1$ edges.
  - In a tree, there is a unique path between any two nodes.
- A **forest** is a graph in which each component is a tree.
- A **star** is a tree where one node (the **center**) is linked to all other nodes.
- A **ring** (or **circle**, or **cycle**) is a connected graph where each node is linked to two others.
- A **bipartite** graph is one that can be partitioned into two sets such that all links connect nodes in "opposite" sets.
  - Buyers&sellers, firms&workers, students&schools, men&women

Network Statistics

Small networks can be visualized directly, but larger networks are harder to visualize and describe.

It's therefore useful to define several **summary statistics** to describe and compare networks (here focusing primarily on undirected graphs):

- Degree distribution (how dense?)
- Diameter and average path length (how tightly connected?)
- Clustering (are friends-of-friends friends?)
- Centrality (which nodes are central or important?)
- Homophily (are nodes of the same "type" more likely to be linked?)

The rest of today's class introduces these summary statistics and discusses some applications.

Neighborhood and Degree

The **neighborhood**, N_i, of node i is the set of nodes to which it is linked: $N_i = \{j : g_{ij} = 1\}$.

For undirected graphs, the **degree**, d_i, of node i is its number of neighbors, or equivalently the cardinality of its neighborhood: $d_i = \sum_j g_{ij} = \sum_j g_{ji} = \#N_i$.

For directed graphs,

- The **out-degree** of node i is $\sum_j g_{ij}$.
- The **in-degree** of node i is $\sum_j g_{ji}$.

One also sometimes seems the terms "out-neighbor" and "in-neighbor".

In applications, if a link from i to j means that i "influences" j, nodes with high out-degree are "influential."
If a link means that i "listens to"[16] or "endorses" j (e.g., hyperlink to j), nodes with high in-degree are influential.

Mean Degree, Density, Sparseness

The **average (mean) degree**, $\bar{d}$, of an undirected network is

$$\bar{d} = \frac{1}{n} \sum_i d_i .$$

Note that if the network has a total of m edges, then we have

$$\sum_i d_i = 2m.$$

Therefore, $\bar{d} = 2m/n$. (Useful equation.)

The **density**, ρ, of an undirected network is the fraction of all possible links that actually exist, given by

$$\rho = \frac{m}{n(n-1)/2} = \frac{\bar{d}}{n-1}.$$

For large networks, often approximate as $\bar{d}/n$.

A network is **sparse** if ρ is small.

- When discussing large networks, this is often taken to mean that $\rho \to 0$ as $n \to \infty$.

Degree Distributions

The **degree distribution**, $P(d)$, of a network describes the proportion of nodes that have different degrees d.

- For a given graph, $P(\cdot)$ is a histogram:
  that is, $P(d)$ is the fraction of nodes with degree d.
- For a random graph model, $P(\cdot)$ is a probability distribution:
  that is, $P(d)$ is the probability that a node has degree d.

A graph is d-**regular** if all nodes have the same degree d
(so $P(d)$ is a degenerate distribution).

- If a graph is d-regular with d odd, it must have an even number of nodes. (Why?)

Degree Distributions (cntd.)

Two types of degree distributions for random graph models:

- $P(d) \leq ce^{-\alpha d}$ for some constants $\alpha > 0$ and $c > 0$: the tails of the distribution fall off exponentially (or faster): large degrees are very unlikely.
- $P(d) = cd^{-\gamma}$ for some constants $\gamma > 0$ and $c > 0$: called a **power-law distribution**, the tails of the distribution are "fat": large degrees are much less unlikely.
  - (Approximate) power laws appear in many settings, including distributions of income, city populations, and internet traffic.
  - Also known as a **scale-free distribution:** a distribution that is unchanged (within a multiplicative factor) under a rescaling of the variable.
  - Appear linear on a log-log plot.

These concepts will play an important role in coming lectures on random graph models.

Diameter and Average Path Length

Let $\ell(i,j)$ denote the distance (shortest path length) between i and j.

The **diameter** of a connected network is the greatest distance between any two nodes:

$$\text{diameter} = \max_{i,j} \ell(i,j)$$

The **average path length** is the average distance between any two nodes:

$$\text{average path length} = \frac{\sum_{i \neq j} \ell(i,j)}{n(n-1)}$$

Average path length is bounded from above by diameter.
In some cases it is much shorter than diameter.

If the network is not connected, one often checks the diameter and the average path length in the largest component.

Clustering

Measures the extent to which my friends are friends with each other.

The simplest such measure is the **overall clustering coefficient** $Cl(g)$, given by

$$Cl(g) = \frac{3 \times \text{number of triangles in the network}}{\text{number of "potential triangles"}},$$

where a "potential triangle" is a triple of distinct nodes (i, j, k) such that $g_{ij} = g_{ik} = 1$.

- Formally,
$$Cl(g) = \frac{\sum_{i;j \neq i;k \neq i,j} g_{ij} g_{ik} g_{jk}}{\sum_{i;j \neq i;k \neq i,j} g_{ij} g_{ik}}.$$
- Note that $0 \leq Cl(g) \leq 1$.
- Also referred to as **network transitivity**: measures extent to which a friend of my friend is also my friend.

Clustering (cntd.)

A different measure of clustering is based on first measuring the "individual clustering" for each node i, then averaging over nodes.

The **individual clustering for node** i is

$$\begin{aligned} Cl_i(g) &= \frac{\text{number of triangles involving } i}{\text{number of potential triangles centered at } i} \\ &= \frac{\sum_{j\neq i;k\neq i,j} g_{ij}g_{ik}g_{jk}}{\sum_{j\neq i;k\neq i,j} g_{ij}g_{ik}} \end{aligned}$$

The **average clustering coefficient** is $Cl^{Avg}(g) = \frac{1}{n}\sum_i Cl_i(g)$.

Consider the undirected "windmill" network, where everyone is linked to the center and one other node.

- Average clustering is close to 1, because $Cl_i(g) = 1$ for everyone except the center.
- Overall clustering is close to 0, because vast majority of potential triangles consist of the center and two individuals who are not linked.

Centrality Measures

There are several measures that capture some notion of the "centrality" or "importance" of a node in a network.

- Different measures capture different notions of centrality, which matter for answering different questions.

- **Degree centrality:** Simply degree divided by $(n-1)$.
- **Closeness/decay centrality:** "On average," how close is the node to other nodes?
  - A simple measure: inverse average distance, or $(n-1) \,/\, \sum_{j \neq i} \ell(i,j)$.
  - A richer measure: **decay centrality**, given by $\sum_{j \neq i} \delta^{\ell(i,j)}$ for some "decay parameter" $\delta \in (0,1)$. (Depends on parameter.)
- **Betweenness centrality:** How important is the node for connecting other nodes?
  - Recall definition from last class:

$$B_k = \sum_{(i,j) \in N : i \neq j, k \neq i,j} \frac{P_k(i,j) \,/\, P(i,j)}{(n-1)(n-2)}$$

Eigenvector-Based Centrality Measures

A more subtle and very important class of centrality measures are based on the self-referential idea that a node is important if it is connected to other important nodes.

- These meaures cannot be computed separately for each node; instead, we compute the measure for all nodes simultaneously via a system of equations.
- These measure are collective called **eigenvector-based centrality measures** (because the calculation involves eigenvectors). They have many applications in this course, including understanding:
  - How Google ranks webpages (PageRank).
  - Which agents in a social network are influential in forming the group's long-run consensus opinion (DeGroot learning).
  - Which firms in a production network are most systemically important (Leontieff input-output analysis).
  - . . . and more.
- We will study this class of measures and its applications next week (start today time permitting).

Homophily and Segregation

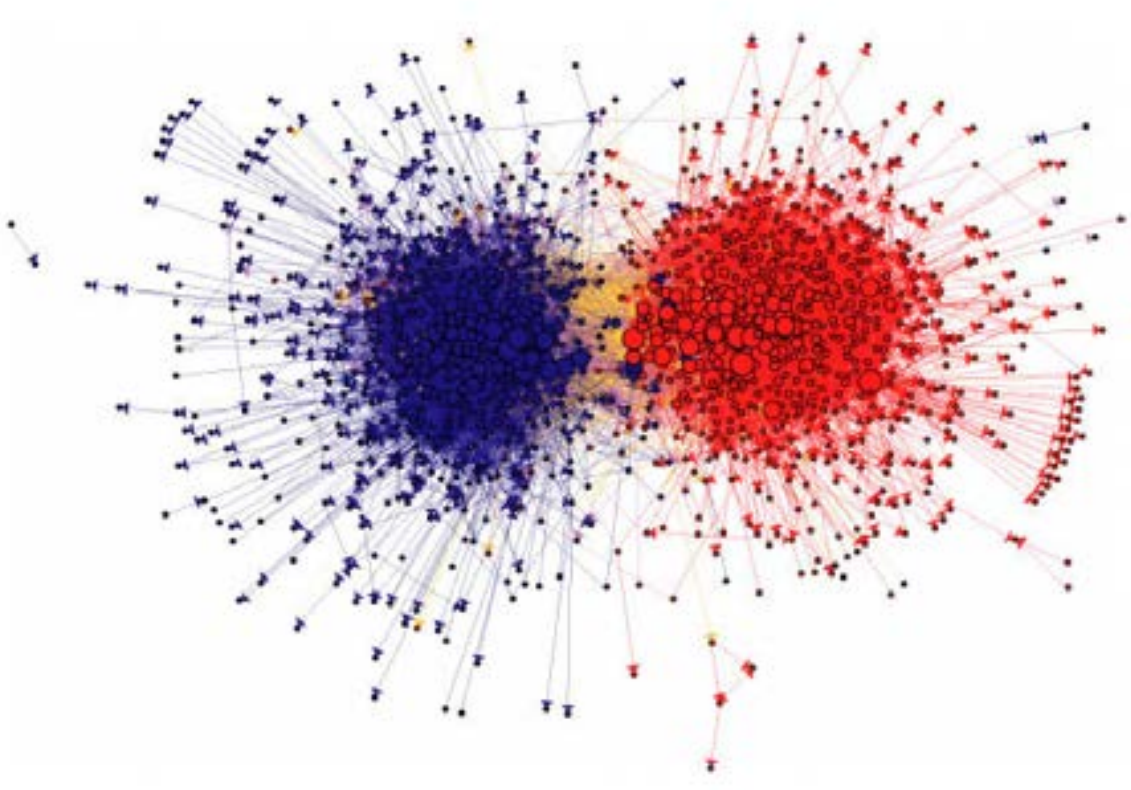
Finally, another kind of network statistic is useful when nodes are of different **types**, or belong to different **groups**.

- Individuals of different gender, race, age, political affiliation, religion, education, etc.
- Liberal vs. conservative blogs (or other media)

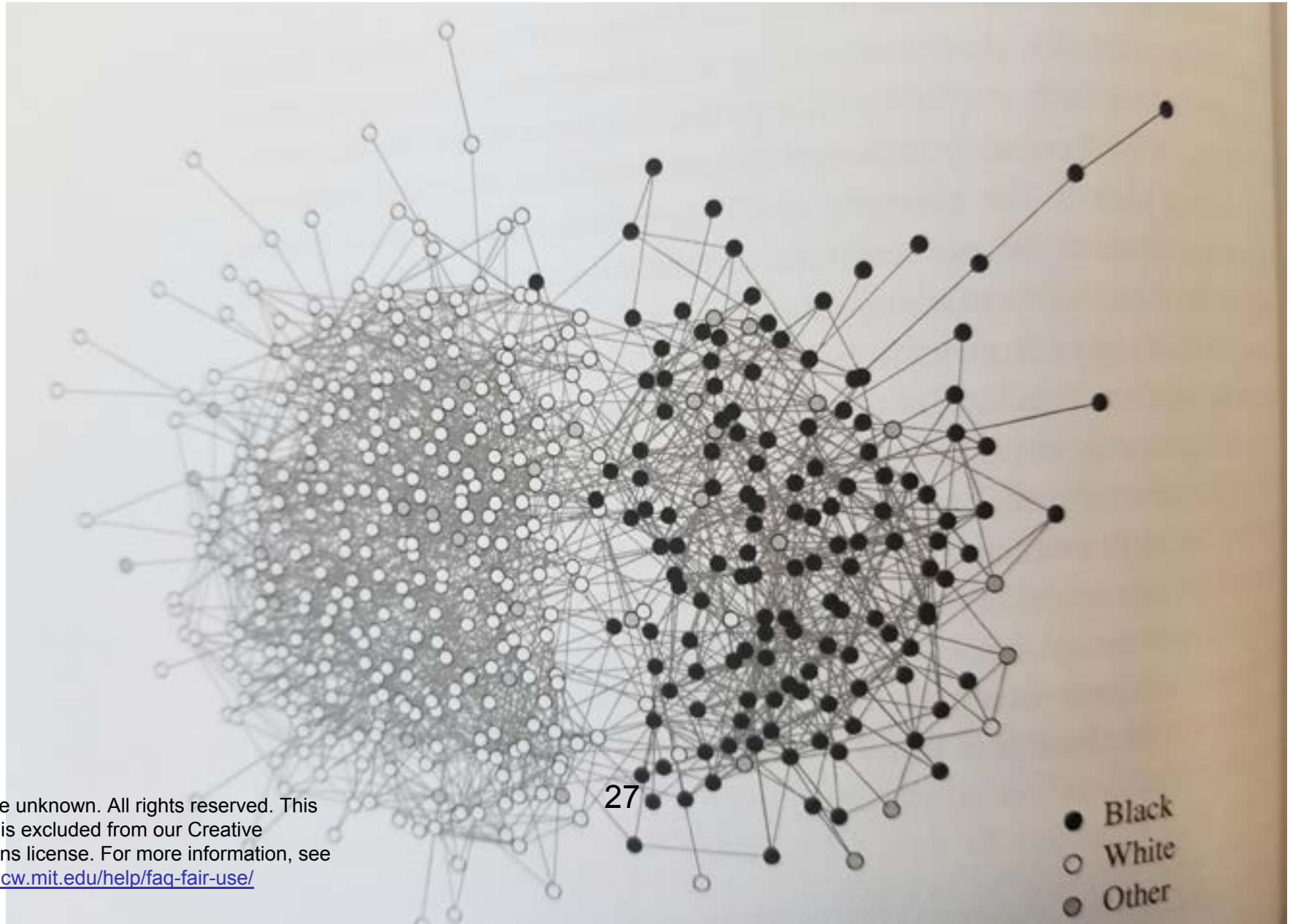
In these settings, a key question is the degree of **homophily**: the extent to which nodes of the same type are more likely to be connected.

- "Similarity begets friendship" —Plato
- "People love those who are like themselves" —Aristotle
- "Birds of a feather flock together" —Proverb

Links Between Political Blogs in the US



The Friendship Network at a US High School



Homophily and Segregation (cntd.)

Issues relating to homophily, assortativeness, and segregation will arise repeatedly in this class.

Later lectures will ask:

- How strong an individual preference of "like for like" (or discrimination of "like against unlike") is needed to result in extreme levels of segregation at the societal level?
- How does homophily affect the speed of diffusion or contagion?
- How does homophily affect whether crowds are wise or foolish? (i.e., whether people successfully aggregate their information, or fall prey to "groupthink" or "echo chambers")

Measuring Homophily

There are different ways of measuring homophily, but the simplest is just to look at the fraction of links that actually exist between individuals of different types, relative to what would be expected if links were formed uniformly at random.

Suppose fraction p_1 of the population is from group 1 and fraction p_2 of the population is from group 2. (May also be other types.)

If links were frandomly distributed, fraction p_1^2 of links would connect two group-1 nodes, and fraction $2p_1p_2$ would connect a group-1 node and a group-2 node.

- If we fix a link and randomly assign the node at each end to type 1 or type 2, we get two type-1's w/ prob p_1^2 and one of each type w/ prob $2p_1p_2$.

Measuring Homophily (cntd.)

Hence, if the fraction of links within group 1 is significantly above p_1^2, this is evidence for homophily (or "assortative matching") within group 1.

If the fraction of links between group 1 and group 2 is significantly below $2p_1p_2$, this is evidence for homophily/assortativity within the groups, or segregation/disassortativity between them.

Introducting Eigenvector Centrality (time permitting)

The simplest measure is **eigenvector centrality**: a non-zero vector $C = (C_i)_{i \in N}$ such that, for some scalar $\lambda > 0$, we have

$$\lambda C_i = \sum_{j \neq i} g_{ji} C_j \quad \text{for all } i \in N.$$

That is, the centrality of each node i is proportional to the weighted sum of the centrality of its neighbors.

- Note that in this definition we have g_{ji} rather than g_{ij}.
- This doesn't matter for undirected graphs, but for directed graphs it says that a node's centrality derives from the centrality of nodes that *point to it*.
- Interpretation: when "important" or "prestigous" nodes point to you, this makes you important/prestigious.
- Equations still hold if multiply C by a scalar. We typically normalize so that $\sum_{i \in N} C_i = 1$.

Eigenvector Centrality (cntd.)

Eigenvector centrality $(C_i)_{i \in N}$ is defined by:

$$\lambda C_i = \sum_{j \neq i} g_{ji} C_j \quad \text{for all } i \in N.$$

It's not immediately obvious whether we can find such a vector C: that is, whether such a measure exists or is unique.

- n linear equations with n unknowns, so looks promising. . .

When is Eigenvector Centrality Well-Defined?

For strongly connected networks, it turns out that eigenvector centrality is always well-defined.

- Recall that a directed network is strongly connected if there exists a directed path between any two nodes.
- In particular, every connected undirected network is strongly connected.
- In general, the network is strongly connected iff for every pair of nodes i, j, there exists a number ℓ such that $\left(g^{\ell}\right)_{ij} > 0$.
  - Matrices g with this property are called **irreducible**.
  - That is, a network is strongly connected if and only if its adjacency matrix is irreducible.

When is Eigenvector Centrality Well-Defined? (cntd.)

In matrix form, the equation for the C_i's is

$$\lambda C = g^T C,$$

where λ is a scalar, C is a $n \times 1$ vector, and g^T is the transpose of the $n \times n$ adjacency matrix (transposed because, for directed graphs, we care about the nodes that link to you, not the nodes you link to).

- That is, C is an eigenvector of g^T, with λ the corresponding eigenvalue.
- The Perron-Frobenius theorem of linear algebra says that, for every irreducible non-negative matrix, its largest eigenvalue is positive, and all the components of the corresponding eigenvector are also positive.
- So, if we let λ be the largest eigenvalue of g^T, the corresponding eigenvector C is non-negative.
- Thus, for any strongly connected network, the eigenvector centrality vector C is well-defined.

Interpretation as Long-Run Population Shares

A useful interpretation of eigenvector centrality as the long-run outcome of a reproduction process (which also explains why it's always well-defined for strongly connected networks):

- Suppose a "virus" starts at a random node in the graph.
- In each period, every virus sends one copy of itself along each link from the node where it is located. Then it dies.
- (So there's 1 virus in period 1, N_i viruses in period 2, $\sum_{j \in N_i} N_j$ viruses in period 3, etc.)
- Letting this process run forever, the virus never dies out (because the network is strongly connected), and we can calculate the long-run fraction of viruses located at each node.
- The long-run fraction of viruses located at node i equals C_i.

(Why? Because the long-run fraction of viruses located at node i is proportional to the long-run fraction of viruses located at nodes that link to node i. This is the relationship that defines eigenvector centrality.)

Perron-Frobenius Theorem

Theorem

For every irreducible non-negative matrix A, its largest eigenvalue r_1 is a positive real number, and the components of the corresponding eigenvector v_1 are also all positive.

The theorem also says more, but this is what we need.

The proof is outside our scope, but we can give an informative informal argument.

Intuition for the Perron-Frobenius Theorem

- Fix any non-negative vector $x(0) \in \mathbb{R}^n$. Suppose that we can write it as a linear combination of the eigenvectors v_i of A:

$$x(0) = \sum_i c_i v_i.$$

- Consider repeatedly multipying $x(0)$ by A. (Matrix multiplication = copying viruses.) After t steps, we get the vector

$$x(t) = A^t x(0) = A^t \sum_i c_i v_i = \sum_i c_i r_i^t v_i = r_1^t \sum_i c_i \left(\frac{r_i}{r_1}\right)^t v_i.$$

- Since r_1 is the largest eigenvalue, $\left(\frac{r_i}{r_1}\right)^t \to 0$ as $t \to \infty$, for all $i \neq 1$. Therefore, $x(t)/r_1^t \to c_1 v_1$. That is, the limiting vector $x(\infty)$ is proportional to the largest eigenvector.
- Since $x(0)$ was non-negative and A is non-negative, each $x(t)$ is also non-negative. Therefore, r_1 must be positive (else oscillates), and every component of v_1 must also be positive.

Other Insights from this Argument

Just like with the viruses, the limiting vector $x(\infty)$ is proportional to the largest eigenvector. This vector defines eigenvector centrality.

We might also ask *how fast* this convergence takes place.

- This is determined by how fast $\left(\frac{r_2}{r_1}\right)^t$ goes to 0 as $t \longrightarrow \infty$ (since $\left(\frac{r_i}{r_1}\right)^t$ goes to 0 faster than this for each $i \geq 3$).
- Bigger gap between first and second eigenvalue $\Longrightarrow$ faster convergence.

14.15 / 6.207 Networks

Spring 2022