6.207/14.15: Networks
Lecture 2: Graph Theory and Social Networks

Daron Acemoglu and Asu Ozdaglar
MIT

September 14, 2009

Outline

- Types of networks
- Graphs: notation and terminology
- Properties of networks:
  - Diameter, average path length, clustering, degree distributions, centrality

Reading:

- Jackson, Chapters 2 and 3
- EK, Chapters 2 and 13

Networks in the Real World

- A network is a set of items (nodes or vertices) connected by edges or links.
- Systems taking the form of networks abound in the world.
- **Types of Networks:**
  - Social and economic networks: A set of people or groups of people with some pattern of contacts or interactions between them.
    - Facebook, friendship networks, business relations between companies, intermarriages between families, labor markets
    - **Questions:** Degree of connectedness, homophily, small-world effects
  - Information networks: Connections of "information" objects.
    - Network of citations between academic papers, World Wide Web (network of Web pages containing information with links from one page to other), semantic (how words or concepts link to each other)
    - **Questions:** Ranking, navigation

Networks in the Real World (Continued)

- **Types of Networks:**
  - Technological networks: Designed typically for distribution of a commodity or service.
    - Infrastructure networks: e.g., Internet (connections of routers or administrative domains), power grid, transportation networks (road, rail, airline, mail)
    - Temporary networks: e.g., ad hoc communication networks, sensor networks, autonomous vehicles
    - **Questions:** Does network structure support performance? Fragility? Cascading failures?
  - Biological networks: A number of biological systems can also be represented as networks.
    - Food web, protein interaction network, network of metabolic pathways

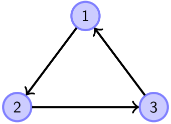
Network Study

- Historical study of networks:
  - Mathematical graph theory: One of the pillars of discrete mathematics
    - Started with Euler's celebrated 1735 solution of the Königsberg bridge problem.
  - Networks also studied extensively in sociology.
    - Typical studies involve circulation of questionnaires, leading to small networks of interactions.
- Recent years witnessed a substantial change in network research.
  - From analysis of single small graphs (10-100 nodes) to statistical properties of large scale networks (million-billion nodes).
  - Motivated by availability of computers and computer networks that allow us to gather and analyze large scale data.
- New Analytical Approach:
  - Find statistical properties that characterize the structure of these networks and ways to measure them
  - Create models of networks
  - Predict behavior of networks on the basis of measured structural properties and models

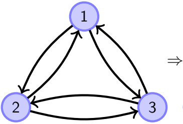
Graphs—1

- We represent a network by a graph (N, g), which consists of a set of nodes $N = \{1, \ldots, n\}$ and an $n \times n$ matrix $g = [g_{ij}]_{i,j \in N}$ (referred to as an adjacency matrix), where $g_{ij} \in \{0, 1\}$ represents the availability of an edge from node i to node j.
  - The edge weight $g_{ij} > 0$ can also take on non-binary values, representing the intensity of the interaction, in which case we refer to (N, g) as a weighted graph.
- We refer to a graph as a directed graph (or digraph) if $g_{ij} \neq g_{ji}$ and an undirected graph if $g_{ij} = g_{ji}$ for all $i, j \in N$.

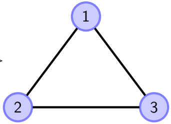
Example 1: $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \Rightarrow$



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



$\Rightarrow$



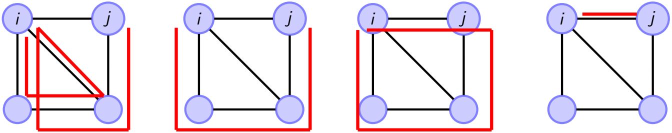
Graphs—2

- Another representation of a graph is given by (N, E), where E is the set of edges in the network.
  - *For directed graphs:* E is the set of "directed" edges, i.e., $(i, j) \in E$.
  - *For undirected graphs:* E is the set of "undirected" edges, i.e., $\{i, j\} \in E$.
- In Example 1, $E_d = \{(1, 2), (2, 3), (3, 1)\}$
- In Example 2, $E_u = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$
- When are directed/undirected graphs applicable?
  - Citation networks: directed
  - Friendship networks: undirected
- We will use the terms network and graph interchangeably.
- We will sometimes use the notation $(i, j) \in g$ (or $\{i, j\} \in g$) to denote $g_{ij} = 1$.

Walks, Paths, and Cycles—1

- We consider "sequences of edges" to capture indirect interactions.
- For an undirected graph (N, g):
  - 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*.
  - A cycle is a path with a final edge to the initial node.
  - A geodesic between nodes i and j is a "shortest path" (i.e., with minimum number of edges) between these nodes.
- A path is a walk where there are no repeated nodes.
- The length of a walk (or a path) is the number of edges on that walk (or path).
- For directed graphs, the same definitions hold with directed edges (in which case we say "a path from node i to node j").

Walks, Paths, and Cycles—2



walk path between i and j cycle shortest path

- *Note:* Under the convention $g_{ii} = 0$, the matrix g^2 tells us number of walks of length 2 between any two nodes:
  - $(g \times g)_{ij} = \sum_k g_{ik} g_{kj}$
  - Similarly, g^k tells us number of walks of length k.

Connectivity and Components

- An undirected graph is connected if every two nodes in the network are connected by some path in the network.
- Components of a graph (or network) are the distinct maximally connected subgraphs.
- A directed graph is
  - connected if the underlying undirected graph is connected (i.e., ignoring the directions of edges).
  - strongly connected if each node can reach every other node by a "directed path".

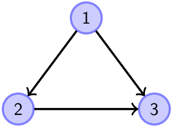
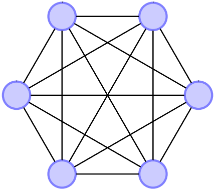


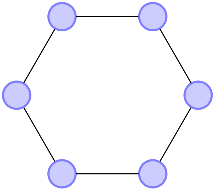
Figure: A directed graph that is connected but not strongly connected

Trees, Stars, Rings, Complete and Bipartite Graphs

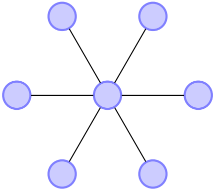
- A tree is a connected (undirected) graph with no cycles.
  - A connected graph is a tree if and only if it has $n - 1$ edges.
  - In a tree, there is a unique path between any two nodes.



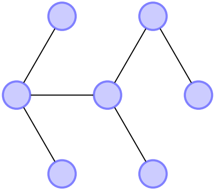
Complete graph



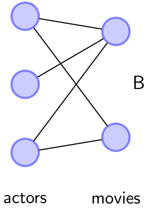
Ring



Star



Tree



Bipartite graph

Neighborhood and Degree of a Node

- The neighborhood of node i is the set of nodes that i is connected to.
- For undirected graphs:
  - The degree of node i is the number of edges that involve i (i.e., cardinality of his neighborhood).
- For directed graphs:
  - Node i's in-degree is $\sum_j g_{ji}$.
  - Node i's out-degree is $\sum_j g_{ij}$.

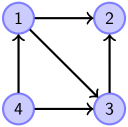


Figure: Node 1 has in-degree 1 and out-degree 2

Properties of Networks

- While a small network can be visualized directly by its graph (N, g), larger networks can be more difficult to envision and describe.
- Therefore, we define a set of **summary statistics** or **quantitative performance measures** to describe and compare networks (*focus on undirected graphs*):
  - Diameter and average path length
  - Clustering
  - Centrality
  - Degree distributions
- A Simple Random Graph Model—Erdös-Renyi model
  - We use the notation $G(n, p)$ to denote the undirected Erdös-Renyi graph.
  - Every edge is formed with probability $p \in (0, 1)$ **independently** of every other edge.
  - Expected degree of a node i is $\mathbb{E}[d_i] =$

Properties of Networks

- While a small network can be visualized directly by its graph (N, g), larger networks can be more difficult to envision and describe.
- Therefore, we define a set of **summary statistics** or **quantitative performance measures** to describe and compare networks (*focus on undirected graphs*):
  - Diameter and average path length
  - Clustering
  - Centrality
  - Degree distributions
- A Simple Random Graph Model—Erdös-Renyi model
  - We use the notation $G(n, p)$ to denote the undirected Erdös-Renyi graph.
  - Every edge is formed with probability $p \in (0, 1)$ **independently** of every other edge.
  - Expected degree of a node i is $\mathbb{E}[d_i] = (n-1)p$

Properties of Networks

- While a small network can be visualized directly by its graph (N, g), larger networks can be more difficult to envision and describe.
- Therefore, we define a set of **summary statistics** or **quantitative performance measures** to describe and compare networks (*focus on undirected graphs*):
    - Diameter and average path length
    - Clustering
    - Centrality
    - Degree distributions
- A Simple Random Graph Model—Erdös-Renyi model
    - We use the notation $G(n, p)$ to denote the undirected Erdös-Renyi graph.
    - Every edge is formed with probability $p \in (0, 1)$ **independently** of every other edge.
    - Expected degree of a node i is $\mathbb{E}[d_i] = (n-1)p$
    - Expected number of edges is $\mathbb{E}[\text{number of edges}] =$

Properties of Networks

- While a small network can be visualized directly by its graph (N, g), larger networks can be more difficult to envision and describe.
- Therefore, we define a set of **summary statistics** or **quantitative performance measures** to describe and compare networks (*focus on undirected graphs*):
  - Diameter and average path length
  - Clustering
  - Centrality
  - Degree distributions
- A Simple Random Graph Model—Erdös-Renyi model
  - We use the notation $G(n, p)$ to denote the undirected Erdös-Renyi graph.
  - Every edge is formed with probability $p \in (0, 1)$ **independently** of every other edge.
  - Expected degree of a node i is $\mathbb{E}[d_i] = (n - 1)p$
  - Expected number of edges is $\mathbb{E}[\text{number of edges}] = \frac{n(n-1)}{2}\, p$

Diameter and Average Path Length

- Let $l(i,j)$ denote the length of the shortest path (or geodesic) between node i and j (or the distance between i and j).
- The diameter of a network is the largest distance between any two nodes in the network:

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

- The average path length is the average distance between any two nodes in the network:

$$\text{average path length} = \frac{\sum_{i \geq j} l(i,j)}{\frac{n(n-1)}{2}}$$

- Average path length is bounded from above by the diameter; in some cases, it can be much shorter than the 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 one another.
- This clustering measure is represented by the overall clustering coefficient $Cl(g)$, given by

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

where a "connected triple" refers to a node with edges to an unordered pair of nodes.

  - Note that $0 \leq Cl(g) \leq 1$.
  - $Cl(g)$ measures the fraction of triples that have their third edge filled in to complete the triangle.
  - Also referred to as network transitivity: measures the extent that a friend of my friend is also my friend.

Clustering (Continued)

- Another measure of clustering is defined on an individual node basis: The individual clustering for a node i is
$$Cl_i(g) = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered at } i}.$$
- The average clustering coefficient is $Cl^{Avg}(g) = \frac{1}{n}\sum_i Cl_i(g)$.

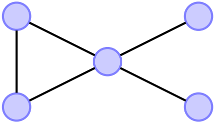


Figure: The overall clustering coefficient for this network is 3/8. The individual clustering for the nodes are 1, 1, 1/6, 0, and 0.

- What is the individual clustering for a node in the Erdös-Renyi model?

Centrality

- A micro measure that captures the importance of a node's position in the network.
- Different measures of centrality
  - Degree centrality: for node i,

$$d_i(g)/n-1, \quad \text{where } d_i(g) \text{ is the degree of node } i$$

  - Closeness centrality: Tracks how close a given node is to any other node: for node i, one such measure is

$$\frac{n-1}{\sum_{j\neq i} l(i,j)}, \quad \text{where } l(i,j) \text{ is the distance between } i \text{ and } j$$

  - Betweenness centrality: Captures how well situated a node is in terms of paths that it lies on (see the Florentine marriages example from the previous lecture).

Degree Distributions

- The degree distribution, $P(d)$, of a network is a description of relative frequencies of nodes that have different degrees d.
  - For a given graph: $P(d)$ is a histogram, i.e., $P(d)$ is the fraction of nodes with degree d.
  - For a random graph model: $P(d)$ is a probability distribution.
- Two types of degree distributions:
  - $P(d) \leq c\, e^{-\alpha d}$, for some $\alpha > 0$ and $c > 0$: The tail of the distribution **falls off faster than an exponential**, i.e., large degrees are unlikely.
  - $P(d) = c\, d^{-\gamma}$, for some $\gamma > 0$ and $c > 0$: Power-law distribution: The tail of the distribution is **fat**, i.e., there tend to be many more nodes with very large degrees.
    - Appear in a wide variety of settings including networks describing incomes, city populations, WWW, and the Internet
    - 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
- What is the degree distribution of the Erdös-Renyi model?