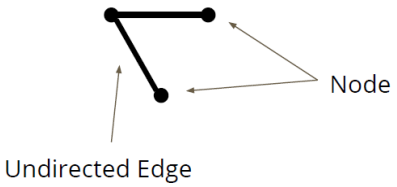
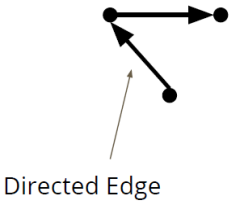
Lecture 17 Network Analysis

Networks are everywhere. To model and analysis these networks, we reuse what is already defined in Math: the ***Graphs***, which consist of nodes / vertices and edges.

Edges can be either ***directed*** (A is connected to B is different from B is connected to A) or ***undirected*** (connections are symmetrical).





Example for directed edges:

"Follow" on social platform. E.g., A is a fan of B and follows B, but B may not follow A.

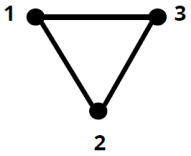
Example for undirected edges:

Friendship that is assumed to be mutual. E.g., A and B both recognize each other as a friend.

The definition of Graph (graph theory):

Formally, a graph G is an ordered pair of sets (V, E) where:
- V is the set of all Nodes / Vertices
- E is the set of all Edges

Let G = (V, E) be undirected where V = {1, 2, 3} and E = {(1,2), (1,3), (2,3)}. What does G look like?



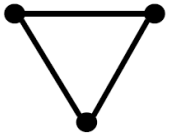
From the computation perspective, how to efficiently store a graph (to make certain query easier)

The answer is ***Adjacency Matrix***, where each column and each row represent a given node and the entry shows whether the 2 nodes are connected. For example, matrix[i][j] = 1, then there is an edge between node i and node j.

*Note that there is no edge between a node and itself, so the diagonal is 0; For undirected graph, the matrix is symmetric.

Adjacency Matrix

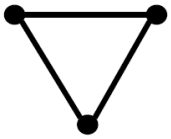
$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$



An alternative way is ***Adjacency List***. For each node, we list all the nodes that are connected to it. It uses less space for storing, but the trade-off is that it runs at a slower speed.

Adjacency List

1 : {2, 3}
2 : {1, 3}
3 : {1, 2}



Graph characteristics:

The **degree** of a Node is the number of edges connected to it.
A **path** between two Nodes is a sequence of edges that joins these two Nodes.
A graph is called **complete** if there is an edge between every pair of Nodes.

*there can be more than 1 path for a pair of nodes; there can be no path between a pair of nodes too. For a complete graph, there is at least 1 path for any pair of nodes.

1. What is the sum of the degrees of all Nodes in a (undirected) Graph as a function of N_E (the number of Edges)?

Because each edge connects 2 nodes and produces a degree of 2, so the total amount of degrees = 2 * N_E

2. How many Edges are in a complete Graph as a function of N_V (the number of Nodes)?

= (new edges connected to node 1) + (new edges connected to node 2) + … + (new edges connected to node N_V)

= $(N_V - 1) + (N_V - 2) + (N_V - 3) + \ldots + 1 + 0$

= $N_V (N_V - 1) / 2$

Network Characteristics

Distribution of edges / node degrees:
- Anomaly detection
- Ranking / Recommendation
- Describe flow through the network

Centrality of a node:
- Identify influencers
- Discover groups / clusterings
- How nodes affect connectivity / flow

Network Analysis

How networks are generated? Through random process, of which the randomness is our model trying to catch.

Let G(N, M) be the set of all graphs with N nodes and M edges. G(N, M) = {G = (V, E) | |V| = N, |E| = M}, where V stands for nodes / vertices and E stands for edges.

Pick uniformly from G(N,M). The probability picking a graph from G(N, M) is:

$$p = \binom{\binom{N}{2}}{M}^{-1}$$

, in which $\binom{N}{2}$ is the total number of possible edges with N nodes and $\binom{\binom{N}{2}}{M}$ is the total number of possible combinations of picking M edges from those edges. p is equal to the inverse of the total numbers of graphs (= the size of G).

Let G(N, p) be generated by randomly connecting nodes with probability p, *independently*. The probability distribution of G(N, p) as a function of a M:

$$f_{G(N,M)} = p^M (1-p)^{\binom{N}{2}-M}$$

, which is a $\frac{N\,(N-1)}{2}$ Bernoulli trial.

Both methods are related in that: G(N,p) conditioned on the event that it has M edges, is equal in distribution to G(N, M).

Proof:

$$P(G(N,p)||E_{G(N,p)}| = M) = \frac{P(G(N,p), |E_{G(N,p)}| = M)}{P(\ |E_{G(N,p)}| = M)}$$
$$= \frac{p^M(1-p)^{\binom{N}{2}-M}}{\binom{\binom{N}{2}}{M}p^M(1-p)^{\binom{N}{2}-M}}$$
$$= \binom{\binom{N}{2}}{M}^{-1}$$

Distribution of the **Degree** of the nodes:

$$P(deg(v) = k) = \binom{N-1}{k}p^k(1-p)^{N-1-k}$$

Note: As N goes to infinity while Np remains constant (i.e. p goes to zero at a comparable rate), the above Binomial distribution converges to a Poisson Distribution

Expected degree = N * p

It is not realistic because it says a high degree node is extremely rare (e.g., someone with a lot of friends on the internet), which is not usually the case in real life. Therefore, a more practical approach to model social networks is the ***Power Law***.

Power Law

Most real-life social networks follow have a degree distribution following a power law of the form

$P(k) = Ck^{-\alpha}$ for some constants C & α

What does this mean?

It is useful to find a simple model / distribution that best describe the network, and different networks may have different models / distributions.

Graph Metrics:

- Diameter

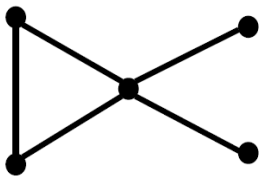
  Let d_{ij} be the shortest path between node i and node j. The diameter of G is defined as

  $$\textbf{Diam(G)} = \max_{ij} d_{ij}$$

  This captures what we refer to as the small world phenomenon.

  *small world phenomenon: everyone in the world can be reached through a short chain of social acquaintances

  Q: What is the Diameter of

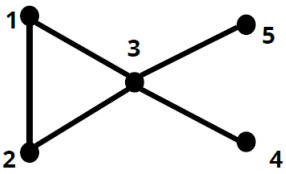
  = 2

- Clustering Coefficient

  C = # triangles / # triplets

  A triangle is a closed triplet. A triplet consists of 3 nodes connected by 2 edges. Triangles and triplets are defined as being centered on a node.

  Ex: What is the clustering coefficient of

  C = (1 + 1 + 1 + 0 + 0) / (1 + 1 + 6 + 0 + 0)
  = 3 / 8

  Node 1, 2 and 3 each has 1 triangle. Node 1 and 2 each has 1 triplet while node 3 has 6 triplets (135, 132, 234, 435, 134, 235). Path direction does not matter.

- Density

  For N nodes and M edges, density = $\frac{2M}{N(N-1)}$. For a complete graph, density is 1.

- Degree Centrality

  The more central a node is, the higher its number of connections.

  = degree of node V

- Closeness Centrality

The more central a node is, the closer it is to all other nodes

$$C_{close}(v) = \frac{1}{\sum_u d(u,v)}$$

- Harmonic Centrality

$$C_h(v) = \sum_{u \neq v} \frac{1}{d(u,v)}$$

Where $d^{-1}(u,v) = 0$ if there is no path between u & v

- Betweenness Centrality

Quantifies the number of times the node acts like a bridge along the shortest path between 2 other nodes

$$C_h(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where $\boldsymbol{\sigma}_{st}$ is the total number of shortest paths from s to t and $\boldsymbol{\sigma}_{st}(v)$ is the number of those shortest paths that go through v