Network Analysis

Boston University CS 506 - Lance Galletti

Example Networks

Internet:
- What will internet traffic through Belgium look like today?
- Anomalous traffic patterns
- Model of the internet

Biology:
- Are certain patterns of interactions among genes more common than expected?
- Which regions of the brain communicate for a given task?

Social:
- Who is friends with whom?
- Who are the influencers?
- What social groups are present?
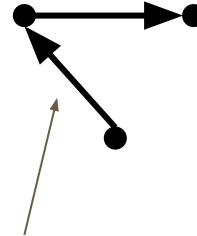- How does information flow through the network?

Graphs

We use Graphs and Graph Theory to model / represent and analyse Networks.

A Graph is comprised of Nodes / Vertices connected by Edges. These Edges can be undirected (where edges are symmetrical connections between Nodes - A is connected to B then B is connected to A) or directed.
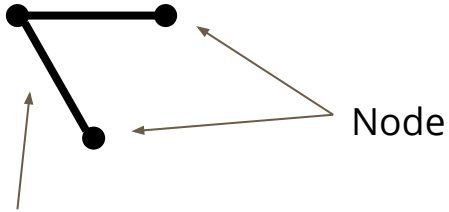
Node

Undirected Edge
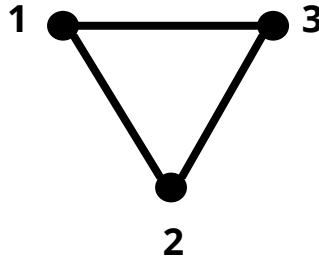
Directed Edge

Graphs

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

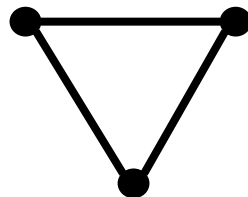Let G = (V, E) be undirected where V = {1, 2, 3} and E = {(1,2), (1,3), (2,3)}. What does G look like?

Graphs

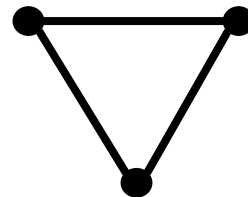How can we efficiently store a graph?

Adjacency Matrix

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

Adjacency List

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

Graphs

Graphs 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.

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

$$2 * N_E$$

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

$$N_V(N_V - 1) / 2$$

Graph Problems

1. Clique Problem (find largest complete subgraph)
2. Coloring Problem (Color a graph with a given a number of colors s.t no two adjacent vertices share a color - or other conditions)
3. Travelling Salesman Problem (Given a list of cities and distances between cities, find the shortest path that goes through all cities and returns to its origin)
4. Shortest Path (given weights on edges, find the shortest path between two nodes)
5. Vertex Cover (When you pick a node, all its adjacent edges get removed. Find the min number of nodes needed to remove all edges from the graph)

Which one do you think is the easiest to solve?

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

Networks / Graphs are generated by processes or functions on its nodes / edges. For example: creating a new account (adding a node), making friends / following / connecting (adding an edge), etc.

The state of a Network / Graph at a given point in time is the **stochastic** result of these processes.

One way we can model the characteristics from the previous slide is by modeling the state of the Graph (i.e. finding the random process that generated the given Graph)

Random Graph Model

1. Let G(N, M) = {G = (V, E) | |V| = N, |E| = M} = the set of all graphs with N nodes and M edges. Pick uniformly from G(N,M).
Ex: G(3,2)   =   { } pick each with probability 1 / 3

In general, what is the probability with which you pick a graph from G(N, M)?

$$p = \left(\binom{\binom{N}{2}}{M} \right)^{-1}$$

2. Let G(N, p) be generated by randomly connecting nodes with probability p, independently. What is the probability distribution of G(N, p) as a function of a number of edges M?
Hint: we are performing N(N-1) / 2 Bernoulli trials inserting edges independently with probability p.

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

Random Graph Model

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:

$$P(G(N,p) \mid |E_{G(N,p)}| = M) = \frac{P(G(N,p), |E_{G(N,p)}| = M)}{P(\,|E_{G(N,p)}| = M)}$$

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

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

Random Graph Model

What is the distribution of the **degree** of 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

Q: What is the expected number of connections / degree for nodes in this Graph?
~ Np

Q: Is this realistic for say social networks?
This means on average we all have similar number of connections and that the probability of a high degree node is exponentially small. Probably not realistic.
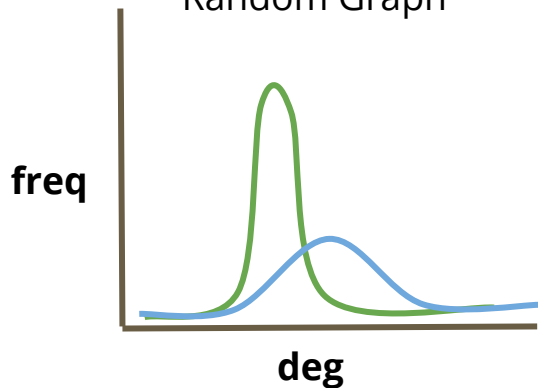
Power Law

Most real-life social networks follow have a degree distribution following a power law of the form
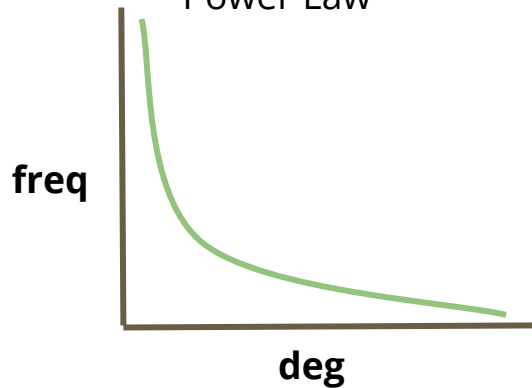
$$P(k) = Ck^{-\alpha} \text{ for some constants C \& } \alpha$$

What does this mean?

Describing / Comparing Graphs

In order to compare graphs, we can define metrics that represent characteristics of the graphs and compare these.

We can talk about metrics that characterize the graph as a whole or characterize a specific node, edge, or set of nodes or edges.

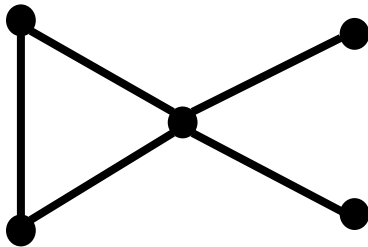# Metrics on Graphs

Diameter

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

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

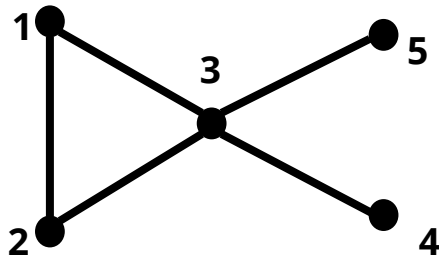This captures what we refer to as the small world phenomenon.
Q: What is the Diameter of

Metrics on Graphs

Clustering Coefficient

$$C = \text{\# triangles} / \text{\# triplets}$$

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

Metrics on Graphs

Density

    Let N = # Nodes, M = # Edges

$$\text{Density} = 2M / N(N-1)$$

Q: what is the density of a complete graph?
A: 1

Metrics on Nodes

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

$$C_{deg}(v) = Deg(v)$$

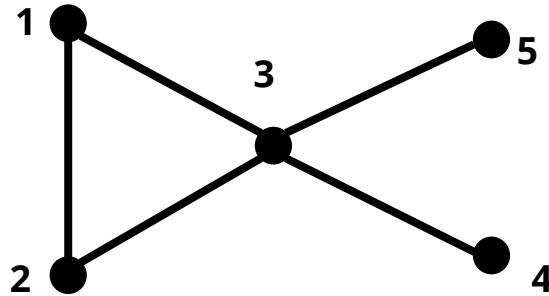**Closeness Centrality**
The more central a node is, the closer it is to all other nodes

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

Metrics on Nodes

For the following Graph

Q: Which node has the highest degree centrality?
A: 3
Q: What is the closeness centrality (where d(u,v) = # edges on the shortest path between u & v) of Node 3?
A: 1/4
Q: What is the closeness centrality of Node 5?
A: 1/7

Metrics on Nodes

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
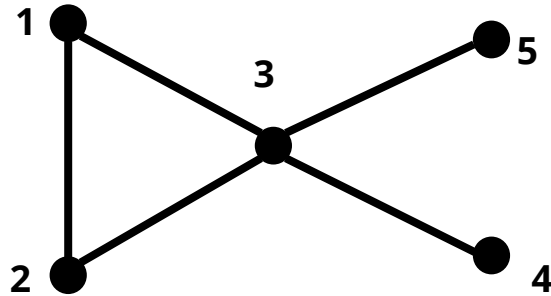
Betweenness Centrality

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

Metrics on Nodes

For the following Graph

Q: What is the harmonic centrality of Node 3?
A: 4
Q: What is the betweenness centrality of Node 3?
A: 10

Recommendation

In homework 3 you will encounter the following problem:

Given a graph and a node v, how can we recommend nodes that v is not connected to?

We can rank all other nodes that v is not connected to from most recommended to least recommended. What scoring function can we use to produce such an ordering?

- Common Neighbors
- Jaccard's Index
- Adamic / Adar Index

Ranking Aggregation

Each of these scoring functions / metrics can generate an ordered list of nodes. We'll refer to this ordered list as a ranking.
Q: How can we compare rankings?

Suppose we have two rankings w_1 , w_2. Can we use a distance function to compare these?

If we have access to the score, we could take the sum of the square differences in scores.
Problem is that different scores are not directly comparable.

Ranking Aggregation

We need a distance function on the ordering itself. Does this remind you of anything?

Disagreement distance?

Let's count the number of inversions / disagreements between the rankings for all pairs of Nodes!

Ranking Aggregation

Kendall τ distance

$d_{\tau}(w_1, w_2)$ = # pairs ranked in different order / # pairs

Node	A	B	C	D
w_1	1	2	3	4
w_2	3	4	1	2

Pair	Agree
AB	✓
AC	✗
AD	✗
BC	✗
BD	✗
CD	✓

$d_{\tau}(w_1, w_2) = 4 / 6$

Ranking Aggregation

Given m Rankings $w_1, \ldots, w_m$, we can generate an aggregate ranking w^*

$$w^* = \arg\max_{w} \sum_{i=1}^{m} d_\tau(w, w_i)$$