VRDI 2019 Networks Breakout
Day 1: Metrics and Measures

Daryl DeFord

June 24, 2019

1 Introduction

The purpose of this session is to introduce you to the types of questions that mathematicians ask about graphs and networks. In particular, we are examining the different types of integer and real valued functions that are used to measure properties of a given graph. One of the interesting features that we will observe is how the questions that we ask distinguish the fields of graph theory and network science from each other, even when we are computing the same metrics.

2 Glossary of Graph Measures

It is useful to establish some notation and definitions so that we are all on the same page.

- We start by introducing some terminology:
  - A graph will be represented $G = (V, E)$ with $|V| = n$ and $|E| = m$
  - We will use nodes or vertices to refer to elements of the set V
  - We will use edges or arcs to refer to elements of the set E
  - Two vertices are neighbors if there is an edge connecting them
  - A vertex and edge are incident if the vertex is one endpoint of the edge
- It is helpful to define some commonly studied types of graphs:
  - A **directed** graph has oriented edges
  - A **multi-** graph allows multiple edges between pairs of nodes
  - A **tree** is a connected graph with no cycles.
  - A **spanning tree** of a given graph is a subset of the edges whose induced subgraph forms a tree.
  - The vertices of a **bipartite graph** can be partitioned into two sets, so that no edges connect vertices that belong to the same set.
  - In a **regular** graph, each vertex has the same number of neighbors.
  - A graph is **Eulerian** if there is a cycle that traverses each edge exactly once.
  - A graph is **Hamiltonian** if there is a cycle that traverses each node exactly once.
  - A graph is **Planar** if it can be embedded in the plane so that none of its edges cross. Given a planar embedding, Euler's formula states that the number of vertices minus the number of edges plus the number of faces is always equal to 2.
- We also want to make use of some functions that are defined on the individual nodes of a given graph:
  - The **degree** of a node is its number of neighbors.

  - The **eccentricity** of a node is the maximum distance between that node and any other node in the graph.

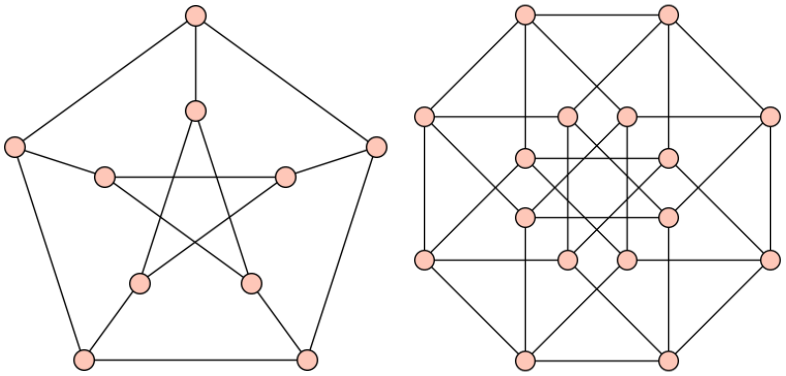
- There are also several common families of graphs that are useful to know:
  - In the **complete graph** K_n, each node is connected to each other node.
  - The nodes of the **path graph** P_n are $\{0, 1, \ldots, n-1\}$ and the edges are $(i, i+1)$ for $0 \leq i \leq n-2$.
  - The nodes of the **cycle graph** C_n are $\{0, 1, \ldots, n-1\}$ and the edges are $(i, i+1 \pmod{n})$ for $0 \leq i \leq n-1$.
  - The nodes of the **star graph** S_n are $\{0, 1, \ldots, n-1\}$ and the edges are $(0, i)$ for $1 \leq i \leq n-1$.
  - The nodes of the **hypercube graph** H_n are all binary strings of length n and there is an edge between any two dtrings that differ in only one place.
  - The **grid graph** $G_{a,b}$ has coordinates (i, j) where $0 \leq i \leq a-1$ and $0 \leq j \leq b-1$ and edges between any nodes

  We next list some common combinatorial measures on graphs:
  - The **order** of a graph is the number of nodes.
  - The **size** of a graph is the number of edges.
  - The **diameter** of a graph is the maximum eccentricity across all nodes.
  - The **radius** of a graph is the minimum eccentricity across all nodes.
  - The **chromatic number** of a graph is the minimum number of colors needed to assign the vertices in such a way that no adjacent nodes are the same color.
  - The **chromatic index** of a graph is the minimum number of colors needed to assign the edges in such a way that no adjacent edges are the same color.
  - The **independence** number of a graph is the maximum size of a set of non-adjacent nodes.
  - The **dominating** number of a graph is the minimum number of a set of vertices so that each vertex is either in the set or adjacent to at least one node in the set.
  - A **vertex cover** is a set of nodes that is incident to each edge of a graph and the size of the minimum vertex cover is the **vertex cover number**.
  - An **edge cover** is a set of edges that is incident to each node of a graph and the size of the minimum edge cover is the **edge cover number**.
  - A **matching** in a graph is a subset of the edges that is incident to each node at most once. A **perfect matching** is incident to each node. The **matching number** is the maximum number of edges in any matching.
  - A **cycle cover** of a graph is a collection of edges so that in the induced subgraph, each node belongs to exactly 1 cycle. Usually, we permit 1 cycles (a node is not attached to any other nodes) and 2 cycles (an edge connects two nodes and they are not connected to any other nodes in the subgraph.

- For many of the graph features above we are interested in enumeration as well as existence. For example, given a specific graph enumerating the following objects is a well–studied problem:
  - The number of spanning trees
  - The number of (perfect) matchings
  - The number of cycle covers
  - The number of Hamiltonian paths
  - The number of extremal everything from the list of combinatorial measures

3 Example: The Petersen Graph

The Petersen graph is a great example of a combinatorial graph. It can be defined in several different ways and is famous among graph theorists for being a counterexample to lots of conjectures.



Let's try to evaluate the Petersen graph[1] G under the metrics introduced in the previous section:

- What is the order of G?
- What is the size of G?
- Is G regular?
- Is G bipartite?
- Is G a tree?
- Draw a spanning tree of G.
- Is G Eulerian?
- Is G Hamiltonian?
- Is G Planar?
- What are the eccentricities of the nodes of G?
- What is the diameter of G?
- What is the radius of G?
- What is the chromatic number of G?
- What is the chromatic index of G?
- What is the independence number of G?
- What is the dominating number of G?
- What is the matching number of G?
- For which lengths t do there exist cycles of length t in G?
- How many maximal matchings does G have?
- How many spanning trees does G have?
- How many Hamiltonian paths does G have?

[1] Or your ego network from week 1

4 Network Measures

Many of the proceeding questions are less well-motivated in an applied setting. For example, trying to determine the chromatic index of a social network is . For more on my thoughts about the differences between networks and graphs, see this. That said there are several common statistics that are measured on both that are simply interpreted differently between the settings.

- (degree) The degree of a node is the number of edges incident to it.
- (degree distribution) The degree distribution of a network is an ordered list of the degrees of each node in the network. The Erdös–Gallai Theorem characterizes when a particular list of integers can be a degree distribution. In network theory the distribution of the degrees is an important invariant of the structure. An interesting facet of complex networks is that while Erdös–Renyi random graphs have a Poisson or binormal degree distribution most observed networks have fat tailed distributions.
  - (power laws) A degree distribution is said to satisfy a power law if the probability of a node in the network having degree k is $k^{-\gamma}$ usually for a value of γ between two and three. These distributions have a scale free property (see 2.2) since multiplication by a constant scales the proportion by the constant to the γ. On a log–log plot these distributions appear linear.
  - (heavy tailed) A long/heavy/fat tailed distribution has more nodes with high degree than a normally/exponentially/uniformly (hopefully clear from context) distributed collection of integers.

The next important class of network statistics deals with paths in networks.

- (geodesics) Shortest paths in a network lead to several important statistics, both for nodes and the entire network. Except in the case of tree networks these geodesics are rarely unique for a given choice of vertices. Many problems dealing with enumerating particular types of paths such as Hamiltonian (non–vertex repeating) or Eulerian (covering every edge) are of significant combinatorial interest. The following statistics are among some of the most frequently studied in complex networks.
  - Node Statistics:
    - The length of the shortest path between node i and j in a network is denoted $\ell_{i,j}$.
    - The eccentricity of a node is $ecc(i) = \max_{i \neq j} \ell_{i,j}$.
    - The closeness centrality of a node is $g_i = \dfrac{1}{\sum_{i \neq j} \ell_{i,j}}$.
    - The number of shortest paths between i and j is $\sigma_{i,j}$.
    - The number of shortest paths between i and j through vertex k is $\sigma_{i,j}(k)$.
  - Network Statistics:
    - The diameter of a graph is $d = \max_{i,j} \ell_{i,j}$.
    - The average shortest path in a graph is $\dfrac{1}{n(n-1)} \sum_{i,j} \ell_{i,j}$.
    - The radius of a graph is $\min_i ecc(i)$.

4.1 Centrality Scores

One of the most common ways to analyze important nodes in a network is through measures of centrality. Broadly these are metrics that attempt to characterize nodes through their importance to the network. The choice of importance metric determines the algebraic computations that are necessary.

- (degree centrality) In this metric the centrality of a degree is set equal to its number of neighbors. This is a simple measure to compute but it does carry some first order information about the network. Citation counts of papers and number of followers on Twitter are examples of this sort of centrality measure.

- (betweeness) Betweeness centrality is a measure of how well connected a node is to the other nodes in a network using the number of paths through the node as a proxy for centrality.
  - Node Statistics
    - The number of edges between neighbors of i is e_i.
    - The clustering coefficient of i is $c(i) = \dfrac{2e_i}{deg(i)(deg(i)-1)}$.
    - The betweeness centrality of i is $b_i = \sum_{k\neq i\neq j} \dfrac{\sigma_{k,j}(i)}{\sigma_{k,j}}$.
    - The load centrality of i is $\dfrac{\sum_{i,j}\sigma_{i,j}(k)}{\sum_{i,j}\sigma_{i,j}}$
  - Network Statistics
    - The clustering coefficient of the network is $\langle c\rangle = \frac{1}{n}\sum_i c(i)$.
    - The δ–clustering coefficient of the network is $c_\delta = \frac{\#\delta}{\binom{n}{3}}$.

  This quantity b_i is measuring how often the node i lies on a geodesic between two arbitrary vertices. For many networks this is a metric with a large range of values that clearly distinguishes between high and low values. The geodesic assumption can be misleading for some natural kinds of network dynamics, such as searching for information. It can also be adjusted to instead reflect the expected number of visits of an absorbing random walk between two arbitrary nodes. Note that this definition is no longer symmetric as can be observed by considering a pendant edge. This methodology has the opposite assumption that there is no optimization in the "flow" across the network.

- (closeness) Closeness centrality is related to the quantity $\ell_i = \frac{1}{n}\sum_j d_{i,j}$, where $d_{i,j}$ is the length of a geodesic from i to j. This has small values when i is close to many nodes so the centrality statistic is defined as the inverse of this number $c_i = \frac{1}{\ell_{i,j}}$. This is a metric that takes a very small range of values across small world type networks, because of the $\log(n)$ diameter assumption and high local clustering.

- (eigenvector) Eigenvector centrality ranks the centrality of nodes in the network by the size of their corresponding entries in the leading eigenvector of the adjacency matrix. Dynamically these processes model random walks on the network. Another technique introduced by Kleinberg is to rate authority centrality and hub centrality separately by computing the eigenvectors corresponding to AA^T and A^TA separately (note that this is only sensible for directed networks). For undirected graphs these are the same since A is symmetric, but for undirected graphs they capture very different information. For examples consider Twitter or the WWW, where having many incoming nodes carries very different information than having many outgoing nodes. Similarly, the property of having arcs to/from hubs (in the power law case) distinguishes between authoritative nodes and standard nodes.

- (Katz) Best technique for directed acyclic graphs where eigenvector centrality cant be measured. The idea is to award each node some measure of centrality at each step to avoid absorbing states in the Markov process. It can be defined as

  $$C_{Katz} = \sum_{k=1}^{\infty}\sum_{j=1}^{n}\alpha^k(A^k)_{i,j}$$

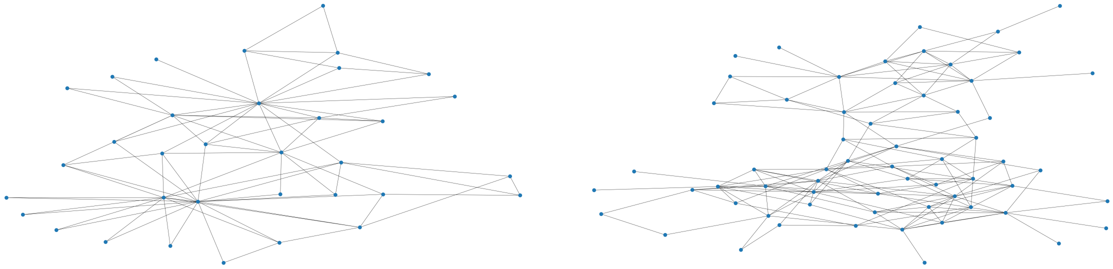
  for some α less than the reciprocal of the largest eigenvalue of A for a single node or

  $$((I-\alpha A^T)^{-1}-I)1$$

  for the entire network. Usually iterative methods are used to calculate these values since it is difficult to prove convergence bounds for this type of system. Google's PageRank algorithm is a version of this centrality measure, with the adjacency matrix normalized to make it stochastic.

5 Examples: Karate and Dolphins

One of the mostly commonly studied networks in the literature is the Zachary karate club - a social network studied by Zachary in this paper. I also really like the dolphin network gathered here. These networks look very different than the graphs considered in the previous sections and so it makes sense that the associated questions would be different.



Let's compute the statistics and centralities for these networks using NetworkX in python. The graph objects are on GitHub. For each of the statistics we want to create two plots: the first showing the distribution of the values and the second highlighting them on the network itself[2]. The python file linked above shows some examples of this. If you save the scores in a vector sorted by node, you can compute the correlations between the scores. How useful would these centralities be on the Petersen graph?

- Degree centrality and distribution
- Eccentricity
- Closeness centrality
- Average shortest path length
- Betweeness centrality
- Load centrality
- Eigenvector centrality
- Katz centrality

Another increasingly common formulation of centrality problems uses a dynamic approach. From a high level, we associate a dynamical operator to the network and score the nodes by their contribution to the dynamics. As an example, consider random walk betweeness centrality, where we replace the shortest paths represented by $\sigma_{i,j}(k)$ with expected number of times that an absorbing random walk beginning at i visits k before reaching j. Here, the dynamics are given by the stochastic matrix associated to the simple random walk.

6 Census Networks

Our final question is to consider how the census networks fit into this picture. Using the json networks from the templates or from mggg-states load the graphs into networkx and evaluate the graph and network statistics that we examined above. Are these census objects graphs or networks? Note that for visulaization purposes you will want to save the positions of the nodes so that they don't have to be recomputed at each step:

```
>>>pos= nx.spring_layout(GA)
>>>nx.draw(GA,pos=pos)
```

[2] For example, coloring the nodes proportional to centrality value or sizing them proportionally to degree.