Lecture 3: Descriptive Statistics

Scribe: Neil Spencer

September 7th, 2016

1 Review

The following terminology is used throughout the lecture

- Network: A real pattern of binary relations
- Graph: $G = (V, E)$ where $E \subset V \times V$. Typically there are no self-edges. Edges are possibly asymmetric (i.e. directed).
- Adjacency matrix: $A_{ij} = 1$ if $(i, j) \in E$, $A_{i,j} = 0$ otherwise.
- Degree of i: The number of edges going to i, given by $\sum_{j=1}^{n} A_{ji}$. If the graph is directed, vertices have an in-degree and an out-degree.
- Connected components: The maximal sets of nodes for which the induced subgraph is connected.
- Counts of subgraphs or motifs (e.g. number of triangles)
- Distance between nodes i and j: the number of edges on the shortest path between i and j (sometimes called the *geodesic* distance).
- Data collection: keep in mind how the data were collected.

2 Graph Summaries

Graph summaries are statistics (i.e. functions of the random graph) which can be used to get a sense of the graph. These statistics are not usually sufficient. Instead, they are used to explore or visualize something about a node, a subgraph, or the graph itself.

2.1 Degree Statistics

One simple statistic of a graph is the degrees of its nodes. They can be summarized as a degree sequence (a count of how many vertices have each degree) or a degree distribution. We can calculate all the traditional summary statistics of the degree distribution such as average degree, the standard deviation of the degree, quantiles, Kurtosis, etc. In networks, the degree distribution is typically right-skewed with a heavy-tail.

The degree of a node is sometimes referred to as its *degree centrality*; high-degree nodes are often more important.

2.2 Distance-based Statistics

Unless stated otherwise, the distance between two nodes is taken to be the geodesic distance. If two nodes are unconnected, the geodesic distance is ∞. Two common distance-based statistics for a graph are:

- Average distance between nodes
- Maximum distance between any two nodes in a graph: this is referred to as the *diameter* of the graph.

Typically, the diameter of a graph is low. Some intuition as to why is given below.

Suppose a graph consists of n nodes. Let $\bar{d}$ denote the average degree of the graph. Given an arbitrary starting node i, let $N(i, j)$ denote the number of nodes which are reachable in a j-step path. Then:
$N(i, 1) \approx \bar{d}$.
$N(i, 2) \approx \bar{d}(\bar{d} - 1)$. Here, the -1 is included to avoid double-counting i.
$\vdots$
$N(i, r) \approx \bar{d}(\bar{d} - 1)^{r-1} \approx (\bar{d} - 1)^r$.
Figure 1 demonstrates the growing of $N(i, r)$ in a network where $\bar{d} = 3$.

The above argument presumes little overlap of a node's neighbourhood that of its neighbours. Given this presumption, the diameter of a graph of n nodes is $O(\log(n))$. This is commonly referred to as the *small-world phenomenon*.

Counterexample: Low dimensional lattices do not exhibit the small-world phenomenon. Instead, p-dimensional lattices have diameters which are $O(n^{1/p})$. Figure 2 demonstrates that a two-dimensional lattice with 25 nodes has diameter 8. This a strong argument that real-world networks

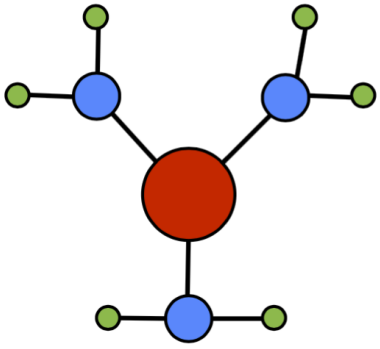

Figure 1: An illustration of $N(i, 0)$ (red node), $N(i, 1)$ (blue nodes), and $N(i, 2)$ (green nodes) for a graph with average degree 3.

should be thought of as living in a very high-dimensional space. For example, each node corresponds to a point in $\mathbb{R}^d$, with nodes connecting if their corresponding points are nearby.

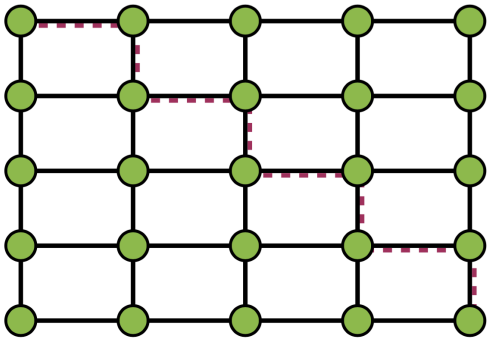

Figure 2: A two-dimensional lattice with 25 nodes. The dotted red line demonstrates the diameter of 8.

Distances provide a tool for defining additional centrality scores, such as *closeness* and *betweenness*.

The closeness centrality of node i (denoted C_i) is n times the reciprocal

of its average distance to other nodes. That is,

$$C_i = \frac{n(n-1)}{\sum_{j \neq i} d_{ij}}.$$

If there are multiple connected components in the graph, there are two alternate definitions:

1. Use only the component of node i (treat it as the whole graph).
2. Use the harmonic mean $C_i = \left(\sum_{j \neq i} d_{ij}^{-1}\right) / (n-1)$.

The betweenness of a node i (denoted B_i) is a measure pf how many shortest paths go through it. Let g_{st} denote the number of geodesic paths linking nodes s and t. Let g_{sti} denote the number of geodesic paths linking nodes s and t which go through i.

$$B_i = \sum_{(s,t) \in E^2} \frac{g_{sti}}{g_{st}}.$$

Figure 3 demonstrates a graph with node i (in red) that has low degree but high betweenness.

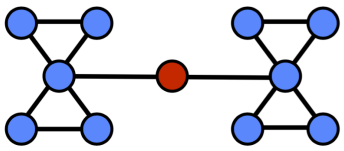

Figure 3: The red node has relatively low degree but high betweenness.

3 Eigenvector Centrality

Eigenvector centrality can be thought of as a circular definition. A node is said to be central if it has links to other central nodes. The eigenvector centrality v_i of node i in a graph can be determined through iterating the following process to equilibrium.

Let $v_j(0) = 1$ for all j. Update the values using the expression

$$v_i(t+1) = \alpha \sum_{j=1}^{n} A_{ij} v_j(t).$$

Under convergence, $\vec{v} = \alpha A \vec{v}$. So $\vec{v}$ must be an eigenvector of αA with eigenvalue 1.

The Perron-Frobenius theorem states that, when A has non-negative entries:

- The biggest eigenvector is positive (all this κ).
- The eigenvector corresponding to κ is non-negative.
- There is one such eigenvector for each connected component.

If w let $\alpha = \kappa^{-1}$, then $\vec{v}$ is the leading eigenvector of αA. Note that if there are multiple connected components, the corresponding eigenvectors are zero for all nodes not included in the component.

The following describes how we know there is a solution to the above problem. It is also a broadly useful technique.

Start with vector $\vec{u}$.

$$\begin{aligned}
\vec{u}(t+1) &= \alpha A \vec{u}(t) \\
&= (\alpha A)^{t+1} \vec{u}(0) \\
&= (\alpha A)^{t+1} \sum_{v_i \text{: eigenvecs}(\alpha A)} \vec{v}_i \left(\vec{v}_i \cdot \vec{u}(0)\right) \\
&= \sum \left(\vec{v}_i \cdot \vec{u}(0)\right) (\alpha A)^{t+1} \lambda_i \vec{v}_i \\
&= \sum \left(\vec{v}_i \cdot \vec{u}(0)\right) \lambda_i^{t+1} \vec{v}_i.
\end{aligned}$$

We know $|\lambda_i| \leq 1$. So this converges to only the summand corresponding to the maximum eigenvalue (1).

Pagerank (of Google fame) is only a slight modification of the above process process.