MATPMD2 : NETWORKS and GRAPH THEORY

8 Graph metrics

When thinking about graphs as networks there are certain questions we might ask.

1. How large is the graph?

   [number of vertices, number of edges, lengths of cycle, lengths of paths, trails, girth, diameter....]

2. How robust/strong is the graph?

   [connectivity, cut vertex, cut edge, cut set, s-t separating set, number of cliques........]

3. How traversible is the graph?

   [paths, trails, cycles, structure, characteristics of particular vertices........]

Let G be a graph with n vertices and write $V = V(G)$ and $E = E(G)$.

8.1 Stepping round a graph

1. Recall a uv path is is a sequence of vertices $v_1, \dots, v_k$ where $v_1 = u$ and $v_k = v$. Such a path has k vertices. Recall that a cycle has no repeated vertices apart from the start/end vertex, otherwise it would not be a cycle. A path P_k has no repeated vertices (and so no repeated edges) and can be considered as a cycle C_k with one edge removed, hence $P_k = C_k - \lambda$. The length of the path is, informally the number of steps taken from one end to the other, and so P_k has length $k - 1$.

2. We could consider the longest path in a graph. If a graph is Hamiltonian it contains a Hamiltonian cycle which is the longest possible cycle in that graph. A longest path can be obtained from a longest cycle by deleting an edge and so in a Hamiltonian graph the longest path is P_n of length $n - 1$.

3. Suppose that G is connected and not a tree. Then G will contain a cycle as a subgraph. The *girth* of a graph is the smallest cycle contained in G.

4. We could consider trails which are walks which allow repeated vertices but not repeated edges. If a graph is Eulerian then it contains an Eulerian trail which uses every edge in the graph precisely once. Consideration of the largest possible Eulerian subgraph might be of use.

5. We could consider the length of the shortest path in a graph. However, unless the graph consists of n isolated vertices, the length of the shortest path will be 1 since between any two adjacent vertices the length of the shortest path between them is 1. Instead we consider the shortest path between any pair of vertices.

8.1.1 Distance

Fix vertices $u, v \in V(G)$. The *graph distance* $d(u, v)$ is the length of the shortest path between vertices u and v. Trivially $d(u, v) = 1$ if $u \sim v$. The graph distance matrix $D = (d_{ij})$ where $d_{ij} = d(v_i, v_j)$.

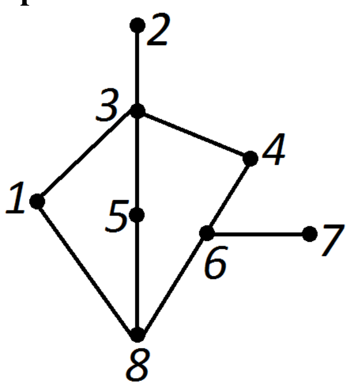
The *diameter* of a graph is the longest shortest path within the graph; the largest value in the distance matrix.

The *average diameter* [or eccentricity] is

$$\frac{1}{n} \sum d_{ij}$$

and can be interpreted as a measure of how quickly something can spread through a network.

Example



Distance matrix

$$D = \begin{matrix} 0 & 2 & 1 & 2 & 2 & 2 & 3 & 1 \\ 2 & 0 & 1 & 2 & 2 & 3 & 4 & 3 \\ 1 & 1 & 0 & 1 & 1 & 2 & 3 & 2 \\ 2 & 2 & 1 & 0 & 2 & 1 & 2 & 2 \\ 2 & 2 & 1 & 2 & 0 & 2 & 3 & 1 \\ 2 & 3 & 2 & 1 & 2 & 0 & 1 & 1 \\ 3 & 4 & 3 & 2 & 3 & 1 & 0 & 2 \\ 1 & 3 & 2 & 2 & 1 & 1 & 2 & 0 \end{matrix}$$

Longest shortest path has length 4 = diameter.

Average diameter $= \frac{1}{n}(4) = \frac{4}{8} = 0.5$.

Girth $=$ size of smallest cycle $= 4$.

The value of the diameter gives an indication of the spread of a graph or network but this can be misleading if the graph contains an unusually long shortest path. To overcome this we consider the *effective diamter* which is defined to be the least graph distance greater then 90% of the graph distances. [List distances in descending order, discard first 10% and the next distance is the effective diameter.]

If G is a weighted graph then the weight of a path is the sum of the edge weights and the distance $d(u, v)$ is defined to be the weight of the uv path of least weight. The *weighted distance matrix* is defined by $W = (w_{ij})$ where $w_{ij} = d(v_i, v_j)$.

8.2 Clustering coefficient

Meet Fred. Fred has lots of friends. We might ask if Fred's friends know one another or not. If Fred has lots of friends that do not know each other this might indicate that he is a bit of a loner with a wide variety of interests which have led to friendships. If many of his friends know each other then we may suppose that Fred is happier as part of a group, or groups. Fred's *clustering coefficient* is a measure of the level of friendship between his friends.
[Note these are merely suppositions but may be useful if we are using a network to disseminate information (for example).]

A relation R is said to be *transitive* on some set if $aRb,\ bRc$ implies $aRc.$ The *equals* relation on a set of numbers is transitive since if $a = b$ and $b = c$ then $a = c.$ The *divides* relation on a set of numbers is transitive since if a divides b and b divides c then a divides $c.$ The relation *is a daughter of* on a set of people is not transitive: Penne is a daughter of Jeanne and Jeanne is a daughter of Nelly does not imply that Penny is a daughter of Nelly. The relation *is friends with* on a set of people may or may not be transitive. Suppose Penne is friends with Fred and Fred is friends with Sol. If Penne and Sol are aso friends then the transitivity property is satisfied for this group of three people. What if within the set of people there are three people who do not know each other? The transitivity property does not fail for this group of strangers since a is not friends with b and b is not friends with c to begin with. If the transitivity property is satisfied for *every* group of three people in the set of people under considertion, then this relation is transitive on this set. If there is one group of three people for which this property fails, then the relation is not transitive on this set.

Construct a graph in the obvious way: vertices represent people, adjacent if these people know each other so that $u \sim v$ means person u knows person v.

Consider vertices u, v, w such that $u \sim v$, $v \sim w$ and consider whether or not $u \sim w$. Here v knows u and w and wonders if u and w know each other. [Are my friends also friends?] If u and w know each other then $u \sim w$ and the transitivity property is satisfied.

Note that if vertices u, v, w are pariwise adjacent - all adjacent to each other - then the subgraph induced on these vertices is a 3-cycle.

Informally the clustering coefficient of a vertex v is a measure of the ratio of number of 3-cycles (triangles, triples) in which v appears to the number of 3-cycles v could appear in if all his friends knew each other.

We fix a vertex v in G and consider H, the subgraph induced by $\eta(v)$, the neighbours of v.
Consider two vertices in H, u and w (both adjacent to u). If $u \sim w$ then u knows w and we have a uw edge, and a $v - u - w - v$ 3-cycle. Clearly every edge in H will contribute one to the count of 3-cycles through v. Thus the number of 3-cycles containing v is the number of edges in the subgraph induced by the neighbours of v.

Now the number of neighbours of v is the degree of v and so H has d_v vertices where d_v is the degree of v. If H is complete then it has all possible edges (all friends know each other) and so the maximum number of 3-cycles which can possibly contain v is $\frac{d_v(d_v-1)}{2}$.
Then the clustering coefficient of v is

$$\frac{\text{the number of edges in the subgraph induced by the neighbours of } v}{\frac{d_v(d_v-1)}{2}}.$$

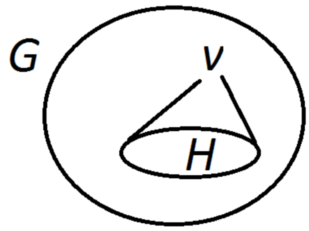
More formally, the clustering coefficient of a vertex v is

$$C_v = \frac{2|\{(u,w) \in E \colon u, w \in \eta(v)\}|}{d_v(d_v-1)}$$

where d_v denotes the degree of v and $\eta(v)$ denotes the neighbourhood of v.

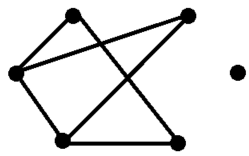
Example

Let G be a graph. Find the clustering coefficient of vertex v in G.



H is subgraph induced by neighbours of v.
Suppose deg(v)=6.

Suppose H is



Here $d_v = 6 = \deg(v)$ and so H is a graph on 6 vertices. We have $\sum_H \deg(v) = 14$ and so H has 7 edges. Then the clustering coefficient of v is

$$C_v = \frac{2|E(H)|}{d_v(d_v-1)} = \frac{2 \times 7}{6 \times 5} = \frac{7}{15}.$$

8.2.1 Cllustering coefficient of the graph.

Note that for a graph with n vertices we will have n clustering coefficients. We take an average of these to obtain the clustering coefficient of the graph G. Thus

$$C_G = \frac{1}{n}\sum_{v \in V} C_v.$$

An alternative (and more common) way of calculating the clustering coefficient of a graph G is

$$C_G = \frac{6 \times \text{ number of triangles in G}}{\text{number of paths length 2}}$$

We can exploit our knowledge of the adjacency matrix A and re-write this as

$$C_G = \frac{tr(A^3)}{\sum_{i \neq j} a_{ij}^{(2)}}$$

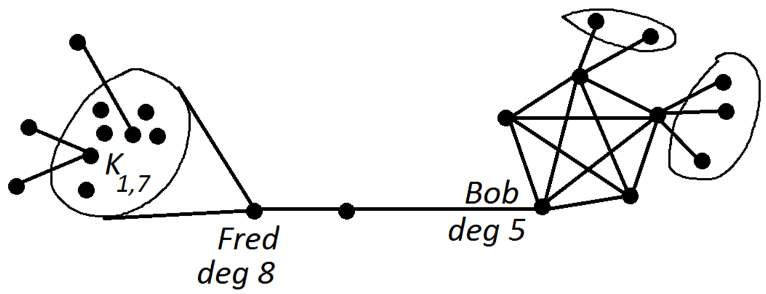
where $tr(A^3)$ denotes trace(A^3). The denominator here is the sum of all entries of A^2 except those on the diagonal so this can be more clearly written as

$$C_G = \frac{tr(A^3)}{\sum_{i,j} a_{ij}^{(2)} - tr(A^2)}$$

8.3 Centrality

Fred is at a convention. He knows lots of people there, and feels that he is an important or *central* piece of the network formed by the people at the convention.

The simplest measure of importance would be to count the number of people he knows at the convention. In graph terms this would be the degree of his vertex and is the *degree centrality* measure. This can be a good indication of importance, or *centrality* but can also be misleading; Fred may know lots of people but if his clustering coefficient is low then it may be that he is not part of the main group. Perhaps he is an independent researcher who collaborates with colleagues on an individual basis rather than being part of a research group.



Here we see that Fred has higher vertex degree than Bob, but Bob is in a more central part of the graph, Fred is on the outskirts or periphery of the graph.

Another measure of importance could be how close Fred is to everybody else. Suppose you can only ask a favour of someone you know. Fred wants to ask a favour of Tim but he does not know Tim. He does know Joy, but she also does not know Tim. However she does know Sunny, and Sunny knows Tim. Fred can ask Joy, who asks Sunny, who asks Tim. In graph terms, Fred has used a path of length 3 to ask his favour. Intuitively Fred would like to involve as few intermediaries as possible so for every person he wants to ask a favour of, he would search for the shortest path between him and that person. A person who could ask favours of anyone using few intermediaries would be said to have a high measure of *closeness centrality*. Put simply, the closeness centrality measures how close your vertex is to all other vertices.

Continuing with the idea of asking favours, Fred notices that he is often used as an intermediary and he feels this makes him an important piece of the network. Without Fred, the people asking favours would have to find someone else to use as an intermediary, his absence would disrupt the network. In graph terms, a frequently used intermediary appears on many of the shortest paths *between* vertices and a vertex with high *betweenness centrality* is crucial to the smooth running of the network.

Definition A vertex is considered *central* if its distance from other vertices is small.

We consider three measures of centrality: degree, closeness and betweeness. Degree centrality measures the degree of a vertex, the closeness centrality measures how close vertices are to each other, and the betweeness centrality is an indicator of how important a vertex is when traversing the graph or network.

8.3.1 Degree centrality

The degree centrality of a vertex v is the number of neighbours of v. In a simple graph this is the degree of v denoted by d_v. We can normalize this metric to a measure between 0 and 1 by dividing by the maximum possible degree of a vertex: $n - 1$. Thus the normalized degree centrality is

$$\frac{d_v}{n-1}.$$

This measure of degree centrality measures the potential of a vertex by taking the ratio of the actual vertex degree to the maximum possible degree.

Another way to measure degree centrality of a vertex v would be to consider comparing the degree of v with the maximum degree of the graph and this would be calculated as

$$\frac{d_v}{d_G}$$

where $d_G = \max\{d_v \colon v \in V(G)\}$.

For a directed graph we would consider in-degrees and out-degrees so the in-degree metric for a vertex would be the number of directed edges going into that vertex and the out-degree metric would be the number of edges leaving that vertex.

We can also consider the idea of degree centrality for a graph G rather than individual vertices.
First consider the graph $K_{1,n-1}$. With the single isolated vertex placed in the middle and the other independent set arranged evenly around it, this would look like a star, and it is commonly called a star graph. It is the graph with highest degree centrality.

Freeman's formula. This compares degree centrality in a graph G to the degree centrality in the star graph.

Fix vertex v of highest degree in graph G so that $d_v = d_G$. Then the degree centrality of the graph is

$$C_G = \frac{\sum_{u \in V(G)} (d_G - d_u)}{(n-1)(n-2)}.$$

For the numerator we calculate the difference in degrees: $d_G - d_u$ for each vertex u in G then calculate the sum of these differences. The denominator performs the same calculation for the star graph. To see this, let S be a star graph on n vertices and let s be the vertex of highest degree in S. Then s has degree $n-1$. The remaining vertices each have degree 1 and so for each of these vertices the difference in degrees is $(n-1) - 1 = n-2$. There are $n-1$ of these vertices and so the sum of the degree differences is $(n-1)(n-2)$.
.

8.3.2 Closeness centrality.

Fix a vertex v in graph G. The closeness centrality of vertex v is

$$\frac{n-1}{\sum_{u \in V} d(u,v)}.$$

To see why recall that $d(uv)$ is the length of the shortest uv path and so $\sum_{u \in V} d(uv)$ is the total shortest distance you must travel to get to v from any other vertices u in the graph. We can obtain this be taking the row sum from the distance matrix for the row corresponding to vertex v. Other than v there are $n-1$ vertices in the graph and so we divide by $n-1$ to get the average distance of v from other vertices. This gives us the metric

$$\frac{\sum_{u \in V} d(u,v)}{n-1}.$$

However this gives us a low value to vertices which are more central and a high value to vertices which are less central. Consequently we take the reciprocal in order to obtain a metric which gives us a high value for vertices which are more central and a low value to vertices which are less central.

Note
Closeness can also be defined as

$$\frac{1}{\sum_{u \in V} d(u,v)}$$

but the normalized version is arguably more useful. [See metabolic networks, Wiener index.]

8.3.3 Betweenness centrality

The *betweeness centrality* quantifies the notion of how often a vertex appears in the shortest paths of a graph. Intuitively a vertex that appears in many such paths has the potential to disrupt the network or graph. [Consider the effect of an accident at a busy road junction.]

Take two vertices u and w. Let P_{uw} be the set of shortest uw paths with $|P_{uw}| = p_{uw}$. Say we wish to travel from u to w and we pick the shortest path according to a uniform distribution; each path has an equally likely chance of being chosen. Then the probability of choosing a particular path is $\frac{1}{p_{uw}}$.
Let $p_{uw}(v)$ be the number of paths in P_{uw} which go through v. Then the *absolute betweeness* can be defined as

$$A_B = \sum_{u,w \in V \setminus v} \frac{p_{uw}(v)}{p_{uw}}.$$

In order to normalize this we need to consider the maximum number of all possible paths uvw. Consider the star graph with central vertex v, and vertices u, w any other vertex. Then all uw paths must go through v. There are $n - 1$ vertices in a star graph, excluding the central vertex and so there are

$$\binom{n-1}{2} = \frac{(n-1)(n-2)}{2}$$

possible uw paths. Since this is the maximum number of uvw paths in any graph we can normalize the betweenness as follows:

$$B_v = \frac{2\sum_{u,w,\in V \setminus v} \frac{p_{uw}(v)}{p_{uw}}}{n^2 - 3n + 2} = \frac{2A_B}{(n-1)(n-2)}.$$

Suggestions for efficient calculations.

1. First fix vertex v - either chosen or given.
2. Calculate $p_{uw}(v)$ for each (u, w) pair. Note that if $u \sim w$ then the shortest path is edge uw which does not go through v. Thus adjacent pairs will contribute nothing to the count of A_B and can be disregarded.
3. Find $P_{uw}(v)$ for each (u, w) pair under consideration. If $P_{uw}(v) = 0$ then we have no contribution to the count and do not have to find P_{uw} for that pair.
4. Suggest you use a table for your calculations.

Example overleaf

9 And finally a very brief mention of Random Graphs

A whirlwind tour of random graphs: Fan Chung, April 1 2008 USCD.