Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Network Science
Lecture 4.06

EEU45C09 / EEP55C09
Self Organising Technological Networks

Nicola Marchetti
nicola.marchetti@tcd.ie

Betweeness Centrality

A different concept of centrality is betweenness centrality which gives a measure of the extent to which a vertex lies on paths between other vertices.

Betweenness Centrality: Motivation

Suppose in our network, we have some quantity flowing from vertex to vertex. For example in a social network, we may have news, information or rumours. In the internet, we may have data packets.

Let us assume that there is a path between every two points in the network and that messages always take the shortest path (geodesic).

If we wait a suitably long time, on average, how many messages will have passed through each vertex?

Betweenness Centrality: Motivation

If we assume that messages are passing down each geodesic (shortest path) at the same rate, the number passing through each vertex is simply proportional to the number of geodesic paths the vertex lies on.

This number of geodesic paths which a vertex lies on is called the betweenness centrality.

Vertices with high betweenness centrality may have a high degree of influence due to their control over information.

Betweenness Centrality

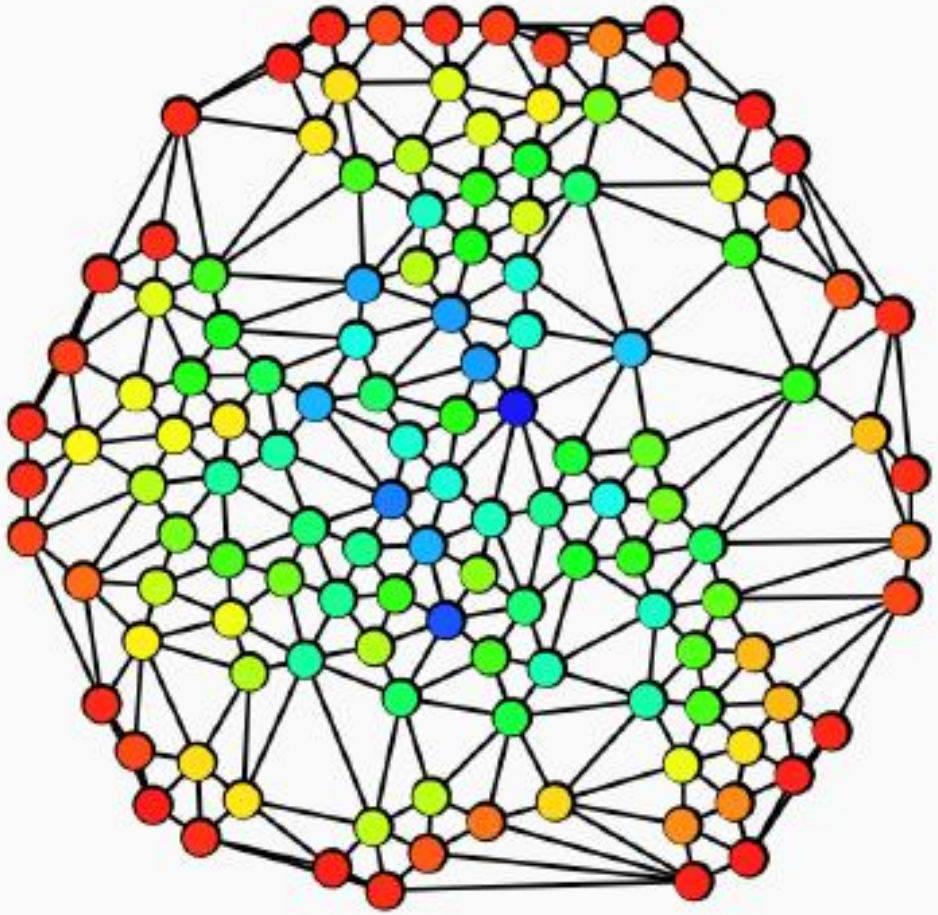


Figure 2: An undirected graph with colour based on Betweenness centrality. The blue nodes have highest BC and red the lowest.

Betweenness Centrality

Vertices with high betweenness centrality in this message-passing scenario are the ones through which the largest number of messages pass.

If the vertices get to read the messages or if they must be paid to pass these messages along, then they can derive a lot of power or influence from their position in the network.

As a corollary to this fact, the high BC vertices are the ones whose removal inflicts the most damage on the network. Communications between other vertices will be highly disrupted since they lie on the path between many pairs of vertices.

In reality, not all vertices communicate through the shortest paths in a network, nonetheless, BC gives us an approximate guide to the influence of a vertex over flow of information in the network.

Betweenness Centrality

We need a more formal definition for the BC.

For the moment, suppose that we are looking at undirected networks in which there is at most one geodesic path between any pairs of vertices (There may be zero paths if the vertices are in different components).

Consider the set of all geodesic paths in this network.

The betweenness centrality of vertex i is defined as the number of those paths that pass through i.

Betweenness Centrality

Let n_{st}^i be 1 if vertex i lies on the geodesic path from s to t and 0 if it does not or if no such path exists.

The BC is then defined as:

$$x_i = \sum_{st} n_{st}^i. \tag{1}$$

Note that this counts the geodesic paths from s to t in either direction between each vertex pair (overcounts).

In practice however, this factor of 2 is often ignored since we are interested only in the relative BC's to put an order on the importance of these vertices. Additionally, Eq. 1 applies immediately to directed networks as well. (We are always looking for methods which generalise well.)

Betweenness Centrality

We must now generalise this method for the case where there is more than one geodesic between each vertex pair. The standard extension here is to give each geodesic path a weight inverse to the number of paths.

If there are two geodesic between vertex pair st, each path gets a weight of $1/2$.

The geodesics between a pair of vertices st need not be vertex-independent (they may pass through some of the same vertices along the way). If two or more paths pass through the same vertex then the betweenness sum includes the contribution from each.

Finally, if there are, for example, three geodesic paths between a given pair of vertices and two of them pass through a particular vertex, they contribute $2/3$ to that vertex's betweenness.

Betweenness Centrality

Formally, can express this betweenness for a general network by redefining n_{st}^i to be the number of geodesic paths from s to t which pass through i. Next, we define g_{st} to be the total number of geodesic paths from s to t. The new definition of the betweenness centrality of vertex i is

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}},$$

where we adopt the convention of $n_{st}^i/g_{st} = 0$ if both the numerator and denominator are zero. **When does this happen?**

This definition is equivalent to our message-passing thought-experiment above. Then, x_i is proportional to the average rate at which traffic passes through i.

Betweenness Centrality

A vertex can have a small degree, be connected to vertices with a small degree, however they still have a large betweenness if they fall "between" many other vertices. (These nodes are often called "brokers".)

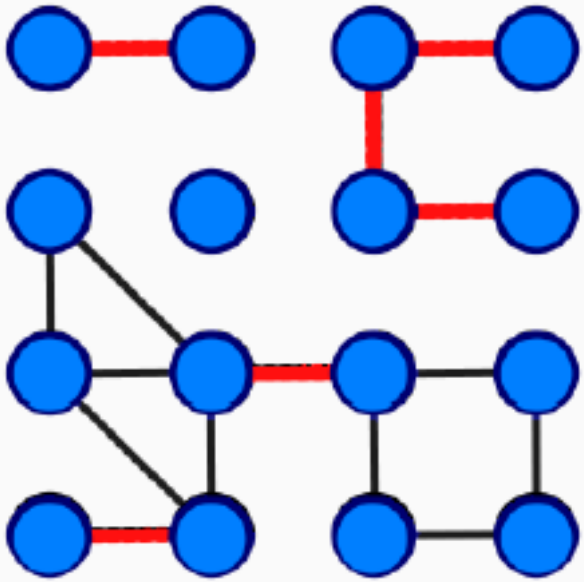
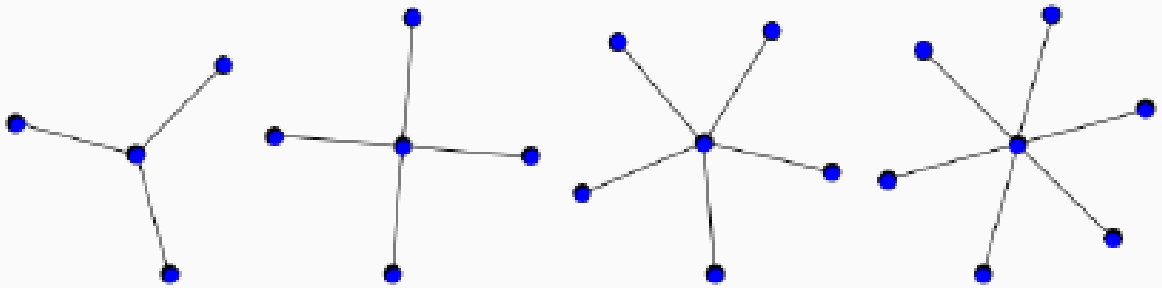


Figure 3: An undirected graph in which the red links are graph bridges.

Betweenness Centrality

As we have done before, we will now look at the range of values over which the BC is typically distributed.

The maximum possible BC occurs for a vertex through which every geodesic path between other vertices passes.



This occurs for the central node in a star graph, where it is incident to the $n-1$ other vertices. By definition, the central vertex is on the graph geodesic path for every other pair of vertices.

Betweenness Centrality

By definition, the central vertex is on all n^2 shortest path between vertices, except for the $n-1$ paths from each radial vertex to itself. The betweenness centrality for the central vertex is therefore,

$$n^2 - n + 1$$

At the other end of the scale, the smallest possible betweenness centrality in a network with a single component is $2n-1$, since at a minimum each vertex lies on a path that starts and ends with itself.

There are $n-1$ paths from a vertex to others and one path from a vertex to itself. Therefore, we have

$$2(n-1) + 1 = 2n - 1.$$

Betweenness Centrality

This lowest BC value occurs when a single vertex is connected to the rest of the network by only a single edge, such as the radial vertices in the star graph.

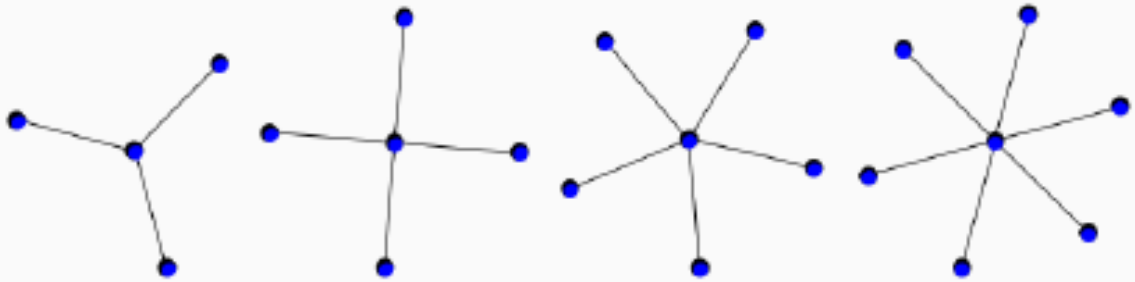


Figure 4: The vertices in the star graph have both largest and smallest possible BC. The central vertex has $x_i = n^2 - n + 1$. The radial vertices have $x_i = 2n - 1$.

Betweenness Centrality

The ratio of the largest and smallest possible betweenness centrality is

$$\frac{n^2 - n + 1}{2n - 1} \simeq \frac{n}{2}.$$

In theory, there could be a factor this large between the largest and smallest BC's, which could become very large for large networks. However, in practice, this is typically smaller. (Few star-like hubs and leaves.)

Betweenness Centrality

The betweenness of actors in the IMDB movie collaboration graph (discussed in the previous section) has also been calculated.

The individual with the largest betweenness centrality in the largest component of the actor network is Spanish actor Fernando Rey (The French Connection 1971). His betweenness score is 7.47×10^8. (The second highest is that of Christopher Lee with 6.46×10^8.)

In contrast, the lowest betweeness centrality is 8.91×10^5. There is a ratio of almost a thousand times, which is a very different story to that of the closeness centrality.

Groups of Vertices

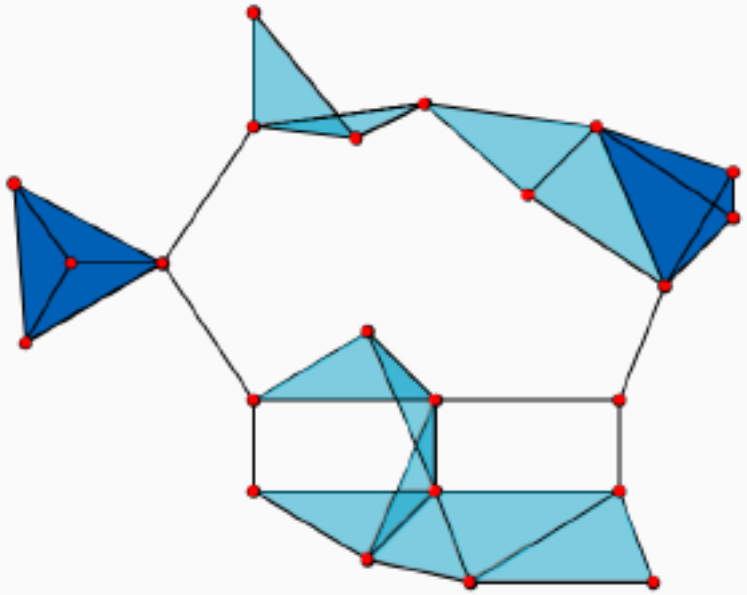
Many networks, including social networks, divide naturally into groups or communities. Networks of people divide into groups of friends and coworkers; the WWW divides into groups of related pages; biochemical networks divide into functional modules.

The definition and analysis of groups of nodes in networks has been well studied. While it may not always be easy to identify independent groups within a network, we often look at the local structure.

The primary constructs of groups of vertices are cliques, plexes, cores and components.

Cliques

A clique is a maximal subset of the vertices in an undirected network such that every member of the set is connected by an edge to every other.



(Maximal means that there is no other vertex in the graph which can be added to the subset while preserving the property that every vertex is connected to every other.)

Cliques

Cliques can overlap: they can share one or more of the same vertices as seen in the last slide.

The occurrence of a clique is an indication of a highly cohesive subgroup within a network.

A set of coworkers in an office or students in a school are likely to form cliques.

$k-$Plex

A clique has the very stringent condition that all member must be connected to all other members.

However, in a social network, it is often the case that a number of people in a large group are not acquainted but in general most people are.

We can relax the stringent clique condition to define a k-plex: a $k-$plex of size n is a maximal subset of n vertices such that each vertex is connected to at least $n - k$ of the other vertices.

If $k = 1$, we recover the definition of a clique. For $k = 2$, then each vertex must be connected to all or all-but-one of the others.

$k-$Core

Closely related to the $k-$Plex is the $k-$Core. A k-core is a maximal subset of vertices such that each is connected to at least k others in the subset. Unlike k-plexes and cliques, k-cores cannot overlap, since by their definition two k-cores that share one or more vertices would just form a single larger k-core.

k-cores are of particular interest for practical reasons since it is very easy to find the set of all k-cores in the network.

Start with the whole network and removing vertices with degree less than k, since by definition these can't be part of a k-core. This may reduce the degree of remaining vertices, so we must repeat the pruning of all vertices with degree less than k.

After the pruning is complete, the graph will now be a k-core or set of k-cores.

Component

A component in an undirected network is a maximal subset of vertices such that each is reachable by some path from each of the others.

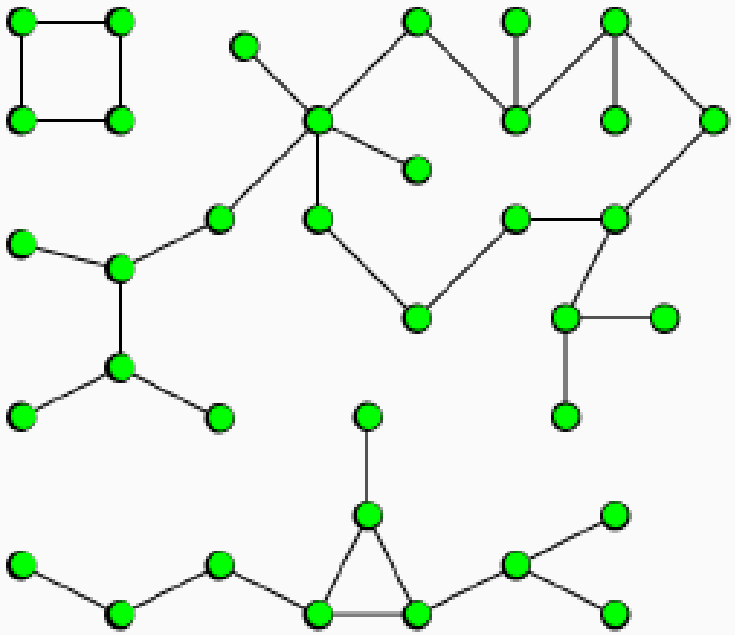


Figure 5: A graph with three different connected components.

k-component

A k-component is a maximal subset of vertices such that each is reachable from each of the other vertices by at least k vertex-independent paths.

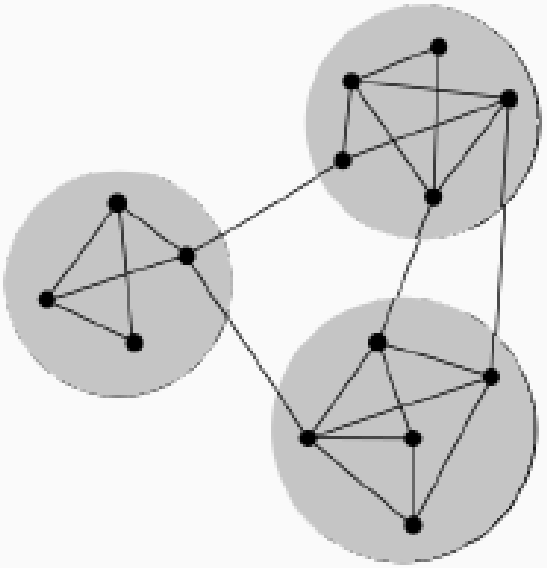


Figure 6: This graph is a 2-component since removing two edges makes it disconnected.

Assortative Mixing (Homophily)

In many networks, nodes have a strong tendency to associate with others which are similar to themselves in some way. This effect is particularly strong in social networks and is called homophily or assortative mixing.

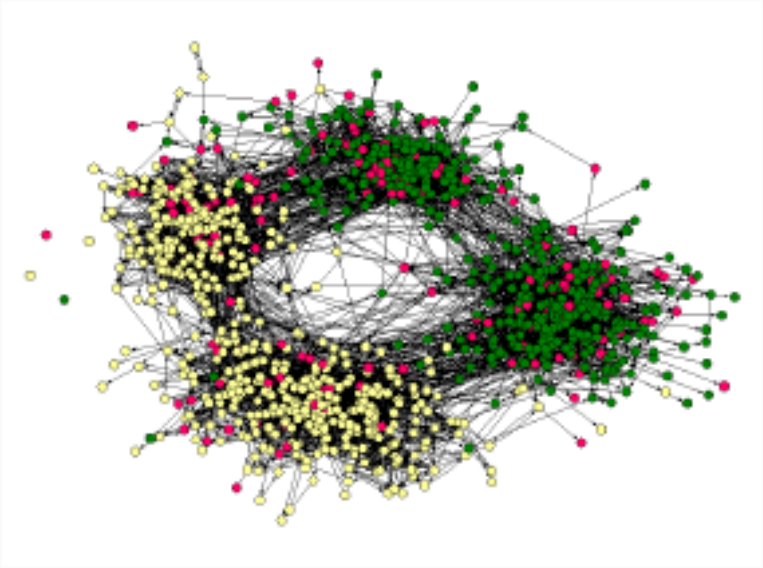


Figure 2: "Race, school integration, and friendship segregation in America," American Journal of Sociology 107, 679-716 (2001).

Assortative Mixing

In a non-social network context homophily is also present:

- Papers in a citation network tend to cite other papers in the same field.
- Web pages written in a particular language tend to link to others in the same language.
- However, most homophily is found in social networks, where assortative mixing occurs between gender, race and nationality etc.

More rarely, disassortative mixing can also occur in network eg. mixing by gender in sexual contact networks.

Assortative Mixing

Suppose we have a network in which the vertices are classified by some characteristic that has a finite set of values eg. hair colour.

The network is assortative if a significant fraction of the edges run between vertices of the same type. (A simple measure of this would be the fraction of edges.)

A more precise way of quantifying this is by finding the fraction of edges that run between vertices of the same type and then, subtract that from the fraction of such edges we expect to find if edges were positioned at random.

Assortative Mixing

For the trivial case, in which all nodes are of the same type, 100% of edges run between nodes of the same type. Though this is the value we expect. The difference between the number of edges and expected number of edges is therefore zero. This value of zero tells us that there is no non-trivial assortativity in the network.

We want some measure which gives us a non-zero value only when the fraction of edges between vertices of the same type is significantly greater that we would expect on the basis of chance.

Assortative Mixing

Let us denote, by c_i, the class or type of vertex i, eg. brown, blonde, red hair. We can assign an integer $1 \ldots n_c$ to these classes, with n_c the total number of classes. The total number of edges which run between vertices of the same type is,

$$\sum_{\text{edges}(i,j)} \delta(c_i, c_j) = \frac{1}{2} \sum_{ij} A_{ij} \delta(c_i, c_j), \tag{1}$$

where $\delta(m, n)$ is the Kronecker delta:

$$\delta(m, n) = \begin{cases} 1 \text{ if } m = n, \\ 0 \text{ otherwise.} \end{cases}$$

Assortative Mixing

Now we calculate the expected number of edges between vertices if edges are placed at random. Consider a particular edge attached the vertex i, which has degree k_i. With m edges in the network, there are $2m$ ends of edges.

The chances that the other end of our edge is one of the k_j ends attached to vertex j is thus $k_j/2m$ if connections are made purely at random. Now, counting all k_i edges attached to i, the total expected number of edges between vertices i and j is then $k_i k_j/2m$.

Assortative Mixing

Therefore, the expected number of edges between all pairs of vertices of the same type is

$$\frac{1}{2}\sum_{ij}\frac{k_i k_j}{2m}\delta(c_i, c_j), \tag{2}$$

where the prefactor of a half prevents double counting again.

Taking the difference of Eq. 1 and Eq. 2 then gives us an expression for the difference between the actual and expected number of edges in the network that join vertices of the same type.

Modularity

The difference gives us,

$$\frac{1}{2}\sum_{ij} A_{ij}\delta(c_i, c_j) - \frac{1}{2}\sum_{ij} \frac{k_i k_j}{2m}\delta(c_i, c_j)$$
$$= \frac{1}{2}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j).$$

Finally, we normalise so that we calculate the fraction of such edges,

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j).$$

Modularity

The quantity Q is called the modularity and is a measure of the extent to which like-vertices are connected to like-vertices in the network. It takes a value less than or equal to one if there are more edges between vertices of the same type than expected in the network and negative if there are less.

Ackowledgement

- Dr Neal McBride, Arista Networks