

# Betweenness-based decomposition methods for social and biological networks.

John W. Pinney\*<sup>1</sup> & David R. Westhead<sup>2</sup>

<sup>1</sup> Faculty of Life Sciences, University of Manchester

<sup>2</sup> Institute of Cellular and Molecular Biology, University of Leeds

## 1 Introduction

This study builds on the work of Newman and Girvan (Newman, 2001; Girvan and Newman, 2002), who introduced a novel method for the decomposition of networks based on the graph theoretical concept of *betweenness centrality*. This method has already been shown to have useful biological applications, in defining subsystems within food webs (Girvan and Newman, 2002) and to investigate biological function in protein-protein interaction networks (Dunn *et al.*, 2005). Here we propose an alternative formulation of betweenness-based decomposition, which allows nodes to participate in more than one module, and which may therefore be more appropriate for the study of social and biological networks.

The betweenness centrality (or ‘betweenness’) of a vertex  $i$  in a simple graph was first defined by Freeman (Freeman, 1977) as the number of shortest paths (also called *geodesics*) between pairs of other vertices that pass through  $i$  (Figure 1a). Where more than one shortest path exists, weight is equally divided between them. For many real-world networks, betweenness can be considered as a measure of the influence of a node over the flow of information between different points in the network.

This definition was generalised by Newman and Girvan to *edge* betweenness, where the betweenness of an edge  $e$  is simply the number of shortest paths between pairs of vertices that run along  $e$  (Figure 1b). Again, where there is more than one shortest path, weight is equally distributed between them.

## 2 Decomposition by edge betweenness

In studying algorithms for identifying communities within simple graphs, particularly social networks such as scientific collaboration networks and friendship networks, Girvan and Newman noted that traditional methods such as hierarchical clustering worked well for tightly-coupled communities, but tended to separate peripheral vertices (with only a few connections) from the clusters to which they seemed naturally to belong. To improve on these existing methods, they proposed a novel algorithm for network decomposition based on edge betweenness:

### **Algorithm BC<sub>e</sub>: edge betweenness decomposition**

For a connected graph  $G$ :

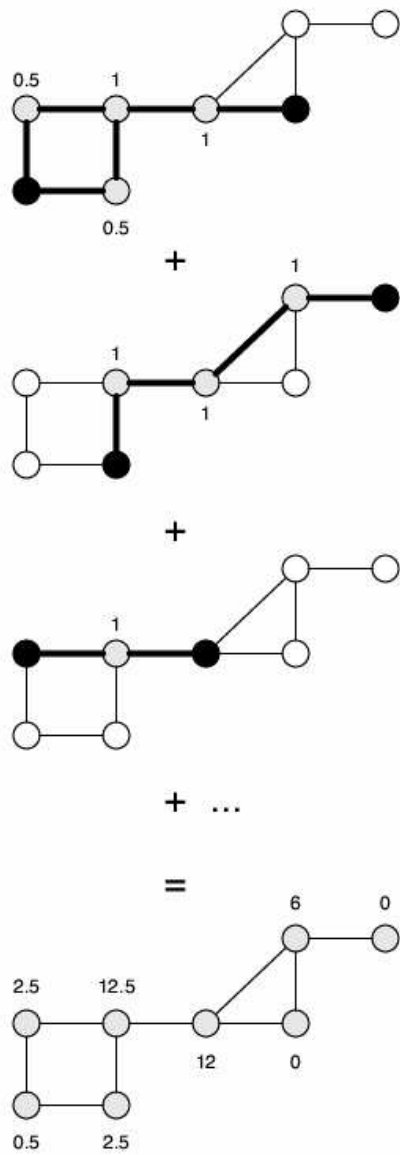
DO

    Calculate betweenness for all edges.

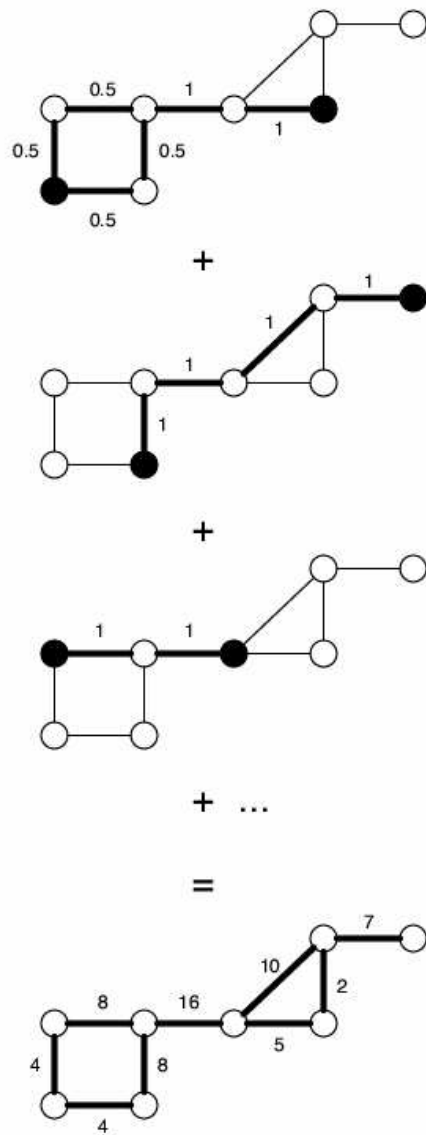
    Remove edge(s) with highest betweenness.

UNTIL  $G$  is no longer connected.

Repeat for each connected component  $G_k$  of  $G$ , until no edges remain.



(a) Vertex Betweenness



(b) Edge Betweenness

Figure 1: The definition of betweenness centrality for vertices (a) and edges (b). Considering a pair of vertices in the graph, vertices (edges) lying on the shortest path between that pair receive a score of 1. Where more than one shortest path exists, vertices (edges) receive a fractional score corresponding to the reciprocal of the number of shortest paths. Summing over all possible pairs of vertices in the graph gives the betweenness centrality for each vertex (edge).

By using betweenness centrality to decompose a graph, Girvan and Newman were able to take into account some of the global character of the networks they studied: edge betweenness is a good way to identify edges that act as ‘bridges’ between two communities, carrying large numbers of shortest paths between pairs of vertices. In contrast, hierarchical clustering methods build up communities starting with purely local information, encapsulated in the form of a weight matrix. The new decomposition method was successful in identifying the community structure in a friendship network from an anthropological study of a karate club, where hierarchical clustering had failed (Girvan and Newman, 2002). The algorithm was also shown to give good results for other social networks such as a college football league and a scientific collaboration network.

### 3 Alternative decomposition algorithms using vertex and edge betweenness

Although Newman and Girvan’s published results are quite impressive, the vertices that are misclassified by their algorithm point towards a potential problem with this type of network decomposition. Because  $BC_e$  creates a network decomposition tree by removing edges, vertices can only be considered to be members of a single community. Therefore, any vertices that participate strongly in more than one community can disrupt the decomposition. In some applications of network decomposition, we are interested in retaining the information about all participants within each community, and we need to allow vertices to appear in more than one cluster. This is particularly true of most biological networks: genes may be transcribed in response to more than one signal, proteins may have more than one biological function, and most metabolites will be involved in more than one metabolic pathway.

To investigate how betweenness centrality may best be employed in the decomposition of networks containing overlapping communities, we have devised an alternative algorithm,  $BC_v$ , based on *vertex* betweenness. With this new algorithm, the network is divided at the vertices with the highest betweenness, which are then copied into the child subnetworks, allowing vertices to be present simultaneously in more than one community.

In most cases, however, the appropriate ‘interfaces’ between subnetworks will include both edges and vertices (see Figure 2). An algorithm combining the removal of edges and vertices therefore seems a good general approach to network decomposition. As a consequence of its definition, edge betweenness will in general tend to be higher than node betweenness. Since there is no obvious way to normalise these two metrics, the edge and vertex betweenness must be compared in some way in order to detect the nature of the interface and decide between edge and vertex removal at each iteration of the algorithm. This is achieved by considering the edge with the highest betweenness, and the betweenness scores of the two vertices that it connects. If the edge acts as a bridge between two communities, then the two vertex scores may be expected to be similar: the majority of enumerated shortest paths that pass through one of them will pass over the bridge and through the other. Conversely, if one of the two vertex scores is much higher than the other then it is clear that the connecting edge is not acting as a bridge. Hence our combined algorithm  $BC_{ve}$  removes the edge with the highest betweenness if its adjacent vertices have similar betweenness scores (e.g. within a 10% tolerance of each other), otherwise it removes the vertex with the highest betweenness, as done in  $BC_v$ .

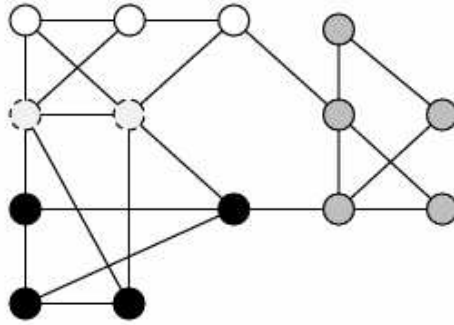


Figure 2: A network including both overlapping and non-overlapping communities. Each vertex has more intra-community than inter-community edges. The dark grey community does not overlap with the others, as it is linked to them by edges only. The black and white communities do overlap, sharing the two nodes marked in light grey (dotted circles).

## 4 Discussion

We have examined how betweenness centrality can be used to decompose networks in different ways, by removing edges, vertices, or both. The edge-removing Girvan-Newman algorithm,  $BC_e$ , is appropriate when vertices are to be separated into discrete groups, but may not be the best choice for certain types of social or biological networks in which vertices are expected to belong to overlapping communities. For this type of network, we propose the  $BC_{ve}$  algorithm as an alternative that allows vertices to belong to more than one module.

These two algorithms have been compared using several network datasets including protein interaction data for the yeast, *S. cerevisiae*, and collaboration networks extracted from the Internet Movie Database ([www.imdb.com](http://www.imdb.com)). Compared to the original Girvan-Newman algorithm,  $BC_{ve}$  has been shown to be better at detecting network community structure where there is significant overlap between clusters.

This algorithm can easily be extended to work with other types of biological network data, including directed graphs (e.g. food webs, gene regulation networks) and bipartite digraphs (e.g. metabolic or signalling networks).

## References

- Dunn, R., Dudbridge, F. and Sanderson, C.M. (2005). The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, **6**, 39.
- Freeman, L.C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, **40**, 35-41.
- Girvan, M. and Newman, M.E. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, **99**, 7821-7826.
- Newman, M.E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **64**, 016132.