Organization Network Analysis & Community Detection of Graphs

Anshuman Guha
Graduate Student
Department of Computer Science
Johns Hopkins University
Baltimore, Maryland, USA

***Abstract*— Understanding communications, information, and decisions flow within a network is important. Current research work focuses on combining graph exploration techniques like organizational network analysis (ONA) & organization structural analysis, graph level diagnostics with traditional community detection algorithms. This would achieve increased operational effectiveness with higher collaboration and exchange of information between the right people, transform organizations to identify formal and informal leaders to facilitate a change and use talent more effectively by minimizing role confusion and redundancy. With the help of organizational network analysis (ONA), a network's exploratory data analysis is performed. Every organization has people (nodes) who serve as critical medium for transfer of ideas and information. Further different subgroups and community detection techniques like clique percolation, label propagation algorithm, k-core decomposition & other methods are used and results are compared.**

Keywords—Organizational Network Analysis, Organizational Structure Analysis, Graph Level Diagnostics, Community Detection, Clustering

INTRODUCTION

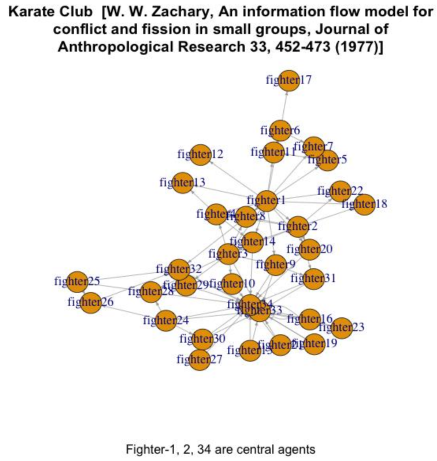
Central nodes share lots of information and influence network effectively. By identifying and managing central nodes, changes can be adopted more quickly and pervasively, and minimize costly disruptions. The organizational structure analysis is done using the measures of connectedness, hierarchy, efficiency, and least upper boundedness. The graph level diagnostics are done using graph diameter, clustering coefficient & graph density metrics. Indicators of centrality identify the most important vertices that influence and provide opinion leadership within an organization. The answer can be given as real-valued function on the vertices of a graph, where the values produced to pass a threshold and provide basis for ranking, which will identify the most important nodes. In this analysis, centrality indices like betweenness centrality, assortativity coefficient, degree, eigenvector centrality, edge density etc. are used to identify important nodes within a network.

I. DATA SET

A. Data Set Information

The Zachary Karate Club is a well-known social network of a university karate club described in "An Information Flow Model for Conflict and Fission in Small Groups" paper by Wayne W. Zachary.

Network description: A social network of a karate club was studied by Wayne W. Zachary for a period of three years from 1970 to 1972. The network captures 34 members of a karate club, documenting 78 pairwise links between members who interacted outside the club. During the study a conflict arose between the administrator "John A" and instructor "Mr. Hi" (pseudonyms), which led to the split of the club into two. Half of the members formed a new club around Mr. Hi, members from the other part found a new instructor or gave up karate. Basing on collected data Zachary assigned correctly all but one member of the club to the groups they actually joined after the split.



Karate Club [W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33, 452-473 (1977)]

Fighter-1, 2, 34 are central agents

Fig 1– Network of the Zachary Karate Club. Node 1 stands for the instructor, node 34 for the president

B. Zachary's methodology

Before the split each side tried to recruit adherents of another party. Thus, communication flow had a special importance and the initial group would likely split at the "borders" of the network. Zachary used the maximum flow – minimum cut Ford–Fulkerson algorithm from "source" Mr. Hi to "sink" John A: the cut closest to Mr. Hi that cuts saturated edges divides the network into the two factions.

II. NETWORK NEASURES OF CENTRALITY

Indicators of centrality identify the most important vertices within this Karate Club dataset. Centrality indices will answer the question "What are the characteristics of important fighters?" The answer can be given as real-valued function on the vertices of a graph, where the values produced to pass a threshold and provide basis for ranking, which will identify the most important nodes.

The word "importance" will suggest different definitions of centrality. "Importance" can be understood as a relation to a type of flow or transfer across the network. This will allow centralities to be classified by the type of flow, which is important for this network. "Importance" can alternately be understood as involvement in the betweenness and cohesiveness of the network. Both of these approaches divide centralities in distinct types.

A. Betweenness Centrality:

In graph theory, betweenness centrality is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.

In this Karate Club, a node with higher betweenness centrality would have more control over the network, because more information will pass through that node. The code logic used to filter high betweenness centrality nodes is:

```
nodes_array <- Initialize array of length as number of nodes
between_centrality = betweenness(adjanceny_matrix)
for ( bc in bet_centrality )
    if ( bc > threshold)
         nodes_array[node]= V(net)$name[node]
  else :  nodes_array[node] = “” #empty string
  Next node
```

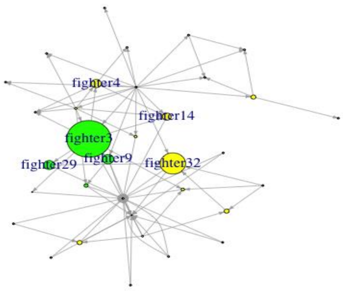


Fig 2– Nodes with High Betweenness Centrality

In this dataset, Fighter-3 is having highest betweenness centrality followed by Fighter-32, 9, 29, 4 & 14. These nodes are of high importance & vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness.

The histogram suggests that Fighter-3 have very high important measure of influence and opinion leadership within this organization.

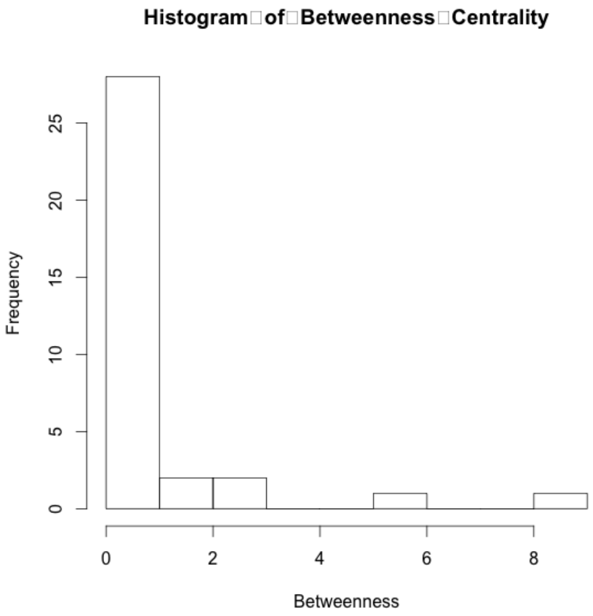


Fig 3– Histogram of Betweenness Centrality

B. Assortativity Coefficient

The assortativity coefficient is the Pearson correlation coefficient of degree between pairs of linked nodes.[2] Positive values of r indicate a correlation between nodes of similar degree, while negative values indicate relationships between nodes of different degree. In general, r lies between −1 and 1. When r = 1, the network is said to have perfect assortative mixing patterns, when r = 0 the network is non-assortative, while at r = −1 the network is completely disassortative.

The assortativity coefficient is given by:

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}.$$

The term q_k is the distribution of the *remaining degree*. This captures the number of edges leaving the node, other than the one that connects the pair. The distribution of this term is derived from the degree distribution.

$$q_k = \frac{(k+1)p_{k+1}}{\sum_{j\geq 1} j p_j}.$$

The assortativity degree for undirected Karate Club network is -0.4844, which suggests that nodes of different degrees are connected to each other.

C. Degree

The degree of a node in a network is the number of connections or edges the node has to other nodes. If a network is directed then nodes have two different degrees. The in-degree, which is the number of incoming edges and the out-degree, which is the number of outgoing edges.

The degree distribution P (k) of a network is then defined to be the fraction of nodes in the network with degree k. Thus if there are n nodes in total in a network and n_k of them have degree k, we have P (k) = n_k/n.

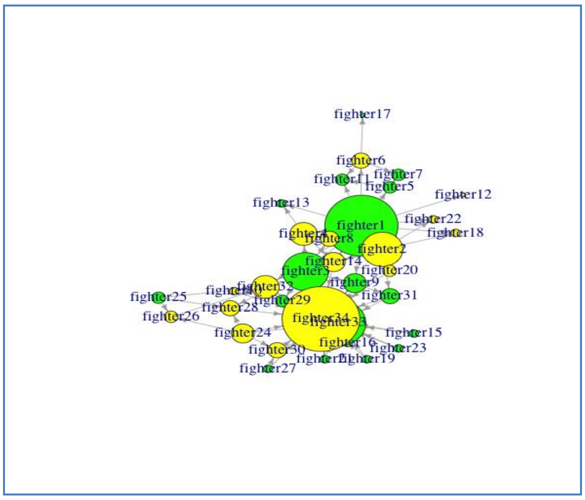


Fig 4 – Degree Plot of all Nodes

In Karate Club, Fighter- 34 & 33 have highest degrees followed by Fighter-1, 2 & 3. Higher degree nodes suggest that these nodes are highly connected to other nodes. Particularly Fighter-3 have high degree & highest betweenness centrality suggesting important & leadership within the club. Here is the histogram for all degrees:

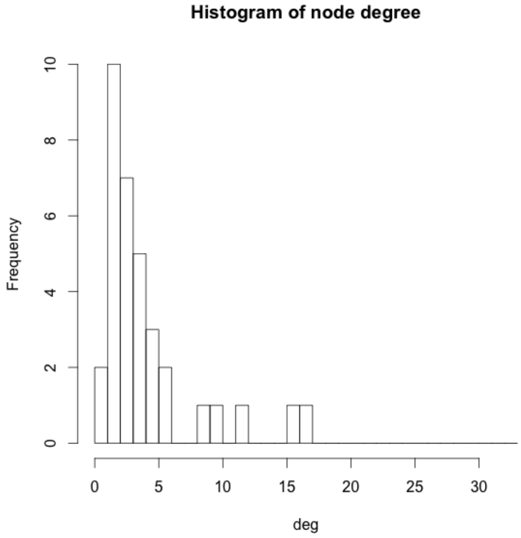


Fig 5– Histogram of Node Degrees

D. *Centrality & Centralization*

Closeness centrality measures the mean distance from a vertex to other vertices. This quantity takes low values for vertices that are separated from others by only a short geodesic distance on average. In a social network a node with lower mean distance to others might find that their opinions reach others in the community more quickly.

The vertex centrality values for all nodes are close to each other with mean value of 0.42178, minimum value of 0.36263 and maximum value of 0.5593. This suggests that mean distances between all nodes are very similar, so information flow across this network is homogenous. The centralization value is 0.2990 and theoretical maximum is 16.24615 will return the theoretical maximum. This low value of centralization might suggest that this network is not organized around its most central points and it is peripheral. But actually this network has central agents spread widely through the graph. The theoretical maximum is absolute deviation (from maximum) conditional on size (which is used by centralization to normalize the observed centralization score).

E. *Eigen Vector Centrality*

In graph theory, eigenvector centrality (also called Eigen centrality) is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. The histogram shown below suggests that few nodes have high eigenvector centrality.

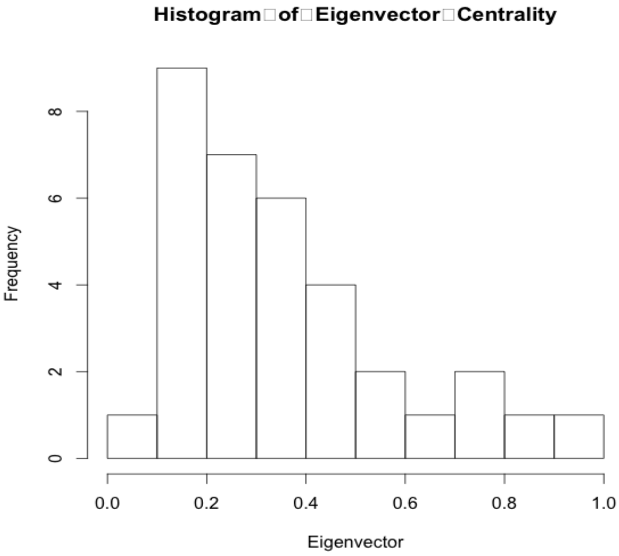


Fig 6– Histogram of Eigenvector Centrality

F. *Edge Density*

Network density describes the portion of the potential connections in a network that are actual connections. A potential connection is a connection that could potentially exist between two nodes – regardless of whether or not it actually does. By contrast, an actual connection is one that actually exists. This person does know that person; this computer is connected to that one.

The edge density for both directed & undirected graphs is 0.06951, which suggests that this network has fewer connections between the nodes as compared to total possible connections.

III. ORGANIZATIONAL STRUCTURAL ANALYSIS

The ideal structure from which to test an organization is the graph theoretical term outtree. (Krackhardt 1994) An outtree is a directed graph where every point except the point at the very “top”, has one and only one edge pointing to it. If these edges represent, say, authority relationships, then one might see how the graph theoretical outtree is more like an archetypical formal hierarchy (Simon 1981) or commonly known as an organization chart. Measures of connectedness, hierarchy, efficiency, and least upper boundedness (LUB) are calculated.

The results of these measures are 1, 0.1372, 0.959, and 0.5 respectively. Each of these four calculated measures gives an analysis of the structure of the organization based on the number of violations that exist in any particular structural arrangement.

A. Connectedness

A violation of connectedness happens when a vertex or node is unable to reach another node within the underlying network. This network shows high connectedness since there are no isolates present. For routine tasks, this measure may not be essential. However, if the task involves research & innovation which requiring consultation and collaboration, then a lack of connectedness could impede the organizations’ ability to adapt.

B. Hierachy

This network has much higher hierarchy, which suggests that asymmetric ties are presented in the reachability directed network derived from the group i.e. there are not too many observed reciprocal ties. This means that a high-level employee can reach a subordinate’s subordinate and opposite is not true i.e. the lower level employee can’t reach his supervisor’s supervisor easily

C. Efficiency

This network has very efficiency value of 0.95. The efficiency is high, because there are very few lateral peer-to-peer ties. Efficiency is a characterization of how dense the network is beyond that, which is absolutely necessary to keep the social group connected to each other. An organization that is trimmed to the point that its network efficiency is maximum also runs the risk of being fragmented because of arbitrary link deletions. On the other hand, low efficiency dense networks require nodes to spend time interacting as opposed to actually doing work.

D. Least Upper Boundedness or LUB

The LUB for this network is 0.5. For a pair of nodes in the network to have a least upper bound (LUB) score they must each have access to a common third person in the network organization to which they can both appeal. Violations of LUB occur when employees have multiple supervisors in the network. High LUB would infer that conflict resolution happens quickly as opposed to organizations that lack high LUB (Doreian 1971). The least upper boundedness is mid-range due to an ambiguity in unity of command.

IV. GRAPH LEVEL DIAGNOSTICS

This set of metrics focuses on network (graph) level properties.

A. Diameter

Diameter indicates the ability of information to quickly move from one side of the network to the other. Diameter for this network is 3. The number of isolates nodes can affect the diameter. Given that an individual’s awareness of resources and knowledge across the network decays somewhere between 2-3 steps through the network, a diameter greater than or equal to 6 will indicate limits to knowledge and resource exchange.

B. Clustering coefficient

Clustering coefficient is similar to density in that it indicates the tendency for nodes to cluster around each other. A higher relative network-clustering coefficient compared to its density can indicate the presence of large clusters in the network. The clustering coefficient for this network is moderate with value 0.1009. Too high of a clustering coefficient leads toward groupthink and lacks the diversity for an agile organization. A low clustering coefficient is good for lean organizations, but poor for agile ones. A moderate clustering coefficient is advisable for agile organizations striving for innovation.

C. Density

Density value for this network is 0.0686 is low, which takes into account the number of isolates present. The density in this network is very moderate. Density also scales poorly with network size. It also has the mathematical equivalence to average degree. Average degree is equal to the density times the network size. Networks with high average degree can become problematic in that nodes are too busy maintaining network ties to have meaningful interactions or generate useful ideas. Networks with low average degree indicate missed opportunities for knowledge and resource exchange.

V. DISTANCES AND PATHS

In the mathematical field of graph theory, the distance between two vertices in a graph is the number of edges in a shortest path (also called a graph geodesic) connecting them. This is also known as the geodesic distance. Average path length is the mean of the shortest distance between each pair of nodes in the network (in both directions for directed graphs). The mean average path length for this graph is 2.41 for directed and 1.269 for undirected graph.

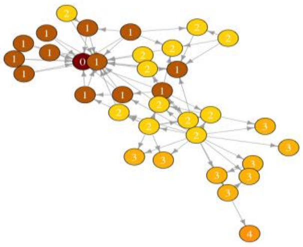


Fig 7– Distance from Fighter-33

VI. SUBGROUP AND COMMUNITIES

A. Cliques

A clique is a sub-set of a network in which the nodes are more closely and intensely tied to one another than they are to other members of the network. In terms of friendship ties, for example, it is not unusual for people in human groups to form cliques on the basis of age, gender, race, ethnicity, religion/ideology, and many other things. The smallest cliques are composed of two nodes: the dyad. A maximal complete sub-graph is such a grouping, expanded to include as many actors as possible.

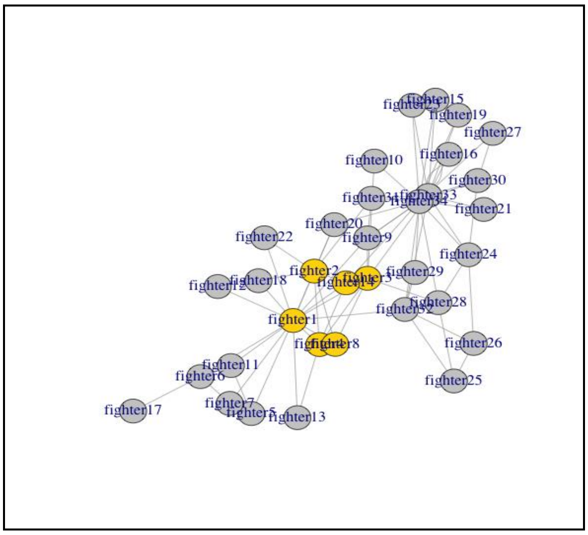


Fig 8 - Clique of fighters-1, 2, 3, 4, 8, 14

B. Community detection

One of the popular methods for community detection based on edge betweenness (Newman-Girvan) where high-betweenness edges are removed sequentially (recalculating at each step) and the best partitioning of the network is selected.

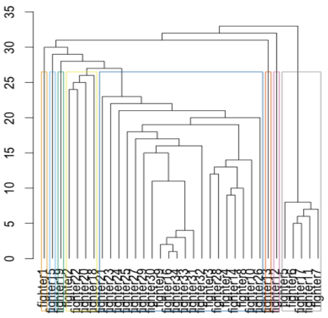


Fig 9 – Hierarchial clustering based on edge betweeness

There are total of eight comunities and modularity value of 0.18828. This high modularity for a partitioning reflects dense connections within communities and sparse connections across communities.

Another method for community detection is based on propagating labels. Node labels are assigned, than replaces each vertex's label with the label that appears most frequently among neighbors. Those steps are repeated until each vertex has the most common label of its neighbors. The below shown community detection graph has 12 subgroups & 0.15 modularity.

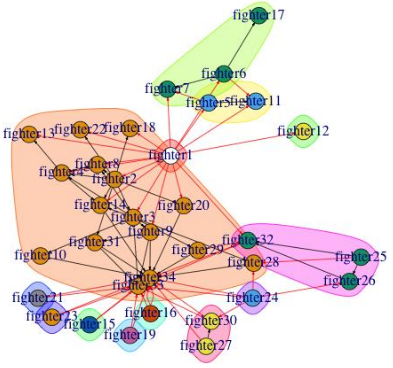


Fig 10 – Community detection using propagating labels

C. K-core decomposition

The k-core is the maximal subgraph in which every node has degree of at least k. The result here gives the coreness of each vertex in the network. A node has coreness D if it belongs to a D-core but not to (D+1)-core. We use the k-core decomposition to visualize large-scale complex networks in two dimensions. This decomposition, based on a recursive pruning of the least connected vertices, allows disentangling the hierarchical structure of networks by progressively focusing on their central cores. By using this strategy we develop a general visualization algorithm that can be used to compare the structural properties of various networks and highlight their hierarchical structure.

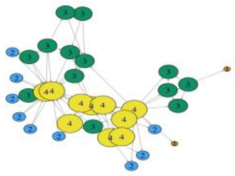


Fig 11 – K-core decomposition shows four subgraphs

D. Hubs and authorities

The hubs and authorities algorithm developed by Jon Kleinberg was initially used to examine web pages. Hubs were expected to contain catalogs with a large number of outgoing links; while authorities would get many incoming links from hubs, presumably because of their high-quality relevant information.

A hub is a component of a network with a high-degree node. Hubs have a significantly larger number of links in comparison with other nodes in the network. The number of links (degrees) for a hub in a scale-free network is much higher than for the biggest node in a random network, keeping the size N of the network and average degree <k> constant. The existence of hubs is the biggest difference between random networks and scale-free networks. In random networks, the degree k is comparable for every node; it is therefore not possible for hubs to emerge. In scale-free networks, a few nodes (hubs) have a high degree k while the other nodes have a small number of links.

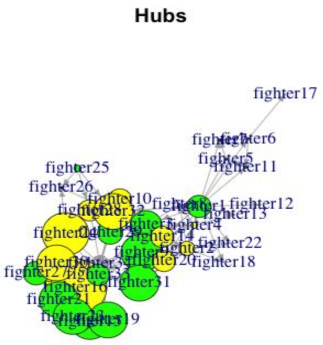


Fig 12– Multiple fighters are hubs in this network

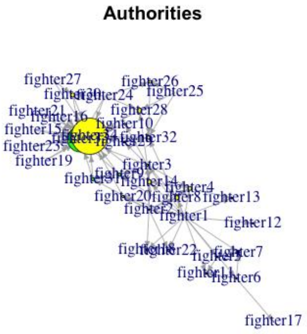


Fig 13– Fighter-32, 33 are main authorities

VII. CONCLUSION

Results obtained suggested that nodes with high degree values are authorities as well in the network with lot of incoming edges. Some of them do not have high betweenness centrality. The vertex centrality values for all nodes are close to each other so information flow across this network is homogenous. Low centralization suggests that network is not organized around its most central points, but it is not peripheral too. Rather this network has central agents spread widely through the graph. The structural analysis revealed that connectedness is high as there are no isolates. The hierarchy is higher and least upper boundedness is in upper-range, which suggests unity of command. This low diameter network has higher relative network clustering coefficient compared to its density, which indicate the presence of large clusters in the network. Different community detection algorithms indicate similar sub-groups within the network & highlight interconnection nodes between these sub-groups. High modularity for a partitioning reflects dense connections within communities and sparse connections across communities.

REFERENCES

[1] Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the national academy of sciences 99.12 (2002): 7821-7826.

[2] Newman, Mark EJ. "Fast algorithm for detecting community structure in networks." Physical review E 69.6 (2004): 066133.

[3] Boccaletti, S., et al. "Detecting complex network modularity by dynamical clustering." Physical Review E 75.4 (2007): 045102.

[4] Mucha, Peter J., et al. "Community structure in time-dependent, multiscale, and multiplex networks." science 328.5980 (2010): 876-878.

[5] Csardi, Gabor, and Tamas Nepusz. "The igraph software package for complex network research." InterJournal, Complex Systems 1695.5 (2006): 1-9.

[6] Zachary, Wayne W. "An information flow model for conflict and fission in small groups." Journal of anthropological research 33.4 (1977): 452-473.

[7] Mucha, Peter J., et al. "Community structure in time-dependent, multiscale, and multiplex networks." science 328.5980 (2010): 876-878.

[8] Dr. Tony Johnson and Dr. Ian McCulloh, “Organizational Network Analysis Using R” August 10, 2016

[9] Christine Sowa, Tony Johnson, and Ian McCulloh, “Organizational Network Analysis Using R-igraph” November 28, 2016