# Maximal Clique Size Versus Centrality: A Correlation Analysis for Complex Real-World Network Graphs

**Natarajan Meghanathan**

**Abstract** The paper presents the results of correlation analysis between node centrality (a computationally lightweight metric) and the maximal clique size (a computationally hard metric) that each node is part of in complex real-world network graphs, ranging from regular random graphs to scale-free graphs. The maximal clique size for a node is the size of the largest clique (number of constituent nodes) the node is part of. The correlation coefficient between the centrality value and the maximal clique size for a node is observed to increase with increase in the spectral radius ratio for node degree (a measure of the variation of node degree in the network). As the real-world networks get increasingly scale-free, the correlation between the centrality value and the maximal clique size increases. The degree-based centrality metrics are observed to be relatively better correlated with the maximal clique size compared to the shortest path-based centrality metrics.

**Keywords** Correlation · Centrality · Maximal clique size · Complex network graphs

## 1 Introduction

Network Science is a fast-growing discipline in academics and industry. It is the science of analyzing and visualizing complex real-world networks using graph theoretic principles. Several metrics are used to analyze the characteristics of the real-world network graphs; among them "centrality" is a commonly used metric. The centrality of a node is a measure of the topological importance of the node with respect to the other nodes in the network [1]. It is purely a link-statistics based measure and not based on any offline information (such as reputation of the node,

N. Meghanathan (✉)
Department of Computer Science, Jackson State University, 18839,
Jackson 39217, MS, USA
e-mail: natarajan.meghanathan@jsums.edu

A. Nagar et al. (eds.), *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, Smart Innovation, Systems and Technologies 44, DOI 10.1007/978-81-322-2529-4_9

cost of the node, etc.). The commonly used centrality metrics are degree centrality, eigenvector centrality, closeness centrality and betweenness centrality. Degree centrality (DegC) of a node is simply the number of immediate neighbors for the node in the network. The eigenvector centrality (EVC) of a node is a measure of the degree of the node as well as the degree of its neighbor nodes. DegC and EVC are hereafter referred to as degree-based centrality metrics. Closeness centrality (ClC) of a node is the inverse of the sum of the shortest path distances of the node to every other node in the network. Betweenness centrality (BWC) of a node is the ratio of the number of shortest paths the node is part of for any source-destination node pair in the network, summed over all possible source-destination pairs that do not involve the particular node. ClC and BWC are hereafter referred to as shortest path-based centrality metrics. Computationally efficient polynomial-time algorithms have been proposed in the literature [1–4] to determine exact values for each of the above centrality metrics; hence, centrality is categorized in this paper as a computationally lightweight metric.

A "clique" is a complete sub graph of a graph (i.e., all the nodes that are part of the sub graph are directly connected to each other). Cliques are used as the basis to identify closely-knit communities in a network as part of studies on homophily and diffusion. Unfortunately, the problem of finding the maximum-sized clique in a graph is an NP-hard problem [3], prompting several exact algorithms and heuristics to be proposed in the literature [5–9]. In this paper, a recently proposed exact algorithm [5] has been chosen to determine the size of the maximum clique for large-scale complex network graphs and extended to determine the size of the maximal clique that a particular node is part of. The maximal clique size for a node is defined as the size of the largest clique (in terms of the number of constituent nodes) the node is part of. Note that the maximal clique for a node need not be the maximum clique for the entire network graph; but, the maximum clique for the entire graph could be the maximal clique for one or more nodes in the network.

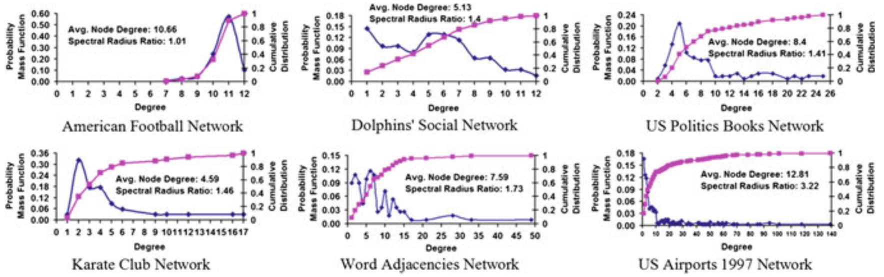
Since the maximal clique size problem is a computationally hard problem and exact algorithms run significantly slower on large network graphs, the paper explores whether the maximal clique size correlates well to one of the commonly studied computationally lightweight metrics, viz., centrality of the vertices, for complex real-world network graphs: if a high positive correlation is observed between maximal clique size and one or more centrality metrics, one could then infer the corresponding centrality values of the vertices as a measure of the maximal clique size of the vertices in real-world network graphs. The work available in the literature so far considers these two metrics separately. This will be the first paper to conduct a correlation study between centrality and maximal clique size for real-world network graphs. To the best of the author's knowledge, there is no other work that has done correlation study between these two metrics (and in general, a computationally hard metric vis-a-vis a computationally lightweight metric) for real-world network graphs.

The rest of the paper is organized as follows: Sect. 2 describes the six real-world network graphs that are used in this paper and presents an analysis of the degree distribution of the vertices in these graphs. Section 3 presents the results of the correlation studies between centrality and maximal clique size at the node level for each of the real-world network graphs. Section 4 concludes the paper. Throughout the paper, the terms 'node' and 'vertex' and 'link' and 'edge' are used interchangeably.

## 2 Real-World Networks and Their Degree Distribution

The network graphs analyzed are briefly described as follows (in the increasing order of the number of vertices): (i) *Zachary's Karate Club*: Social network of friendships (78 edges) between 34 members of a karate club at a US university in the 1970s; (ii) *Dolphins' Social Network*: An undirected social network of frequent associations (159 edges) between 62 dolphins in a community living off Doubtful Sound, New Zealand; (iii) *US Politics Books Network*: Nodes represent a total of 105 books about US politics sold by the online bookseller Amazon.com. A total of 441 edges represent frequent co-purchasing of books by the same buyers, as indicated by the "customers who bought this book also bought these other books" feature on Amazon; (iv) *Word Adjacencies Network*: This is a word co-appearance network representing adjacencies of common adjective and noun in the novel "David Copperfield" by Charles Dickens. A total of 112 nodes represent the most commonly occurring adjectives and nouns in the book. A total of 425 edges connect any pair of words that occur in adjacent position in the text of the book; (v) *American College Football Network*: Network represents the teams that played in the Fall 2000 season of the American Football games and their previous rivalry—nodes (115 nodes) are college teams and there is an edge (613 edges) between two nodes if and only if the corresponding teams have competed against each other earlier; (vi) *US Airports* 1997 *Network*: A network of 332 airports in the United States (as of year 1997) wherein the vertices are the airports and two airports are connected with an edge (a total of 2126 edges) if there is at least one direct flight between them in both the directions. Data for networks (i) through (v) and (vi) can be obtained from http://www-personal.umich.edu/~mejn/netdata/ and http://vlado.fmf.uni-lj.si/pub/networks/pajek/data/gphs.htm respectively.

Figure 1 presents the degree distribution of the vertices in the six network graphs in the form of both the Probability Mass Function (the fraction of the vertices with a particular degree) and the Cumulative Distribution Function (the sum of the fractions of the vertices with degrees less than or equal to a certain value). The average node degree and the spectral radius degree ratio (ratio of the spectral radius and the average node degree) have been also computed; the spectral radius (bounded below by the average node degree and bounded above by the maximum node degree) is the largest Eigenvalue of the adjacency matrix of the network graph, obtained as a result of computing the Eigenvector Centrality of the network graphs. The spectral radius degree ratio is a measure of the variation in the node degree with respect to the average node degree; the closer the ratio is to 1, the smaller the variations in the



American Football Network

Dolphins' Social Network

US Politics Books Network

Karate Club Network

Word Adjacencies Network

US Airports 1997 Network

**Fig. 1** Node degree: probability mass function and cumulative distribution

node degree and the degrees of the vertices are closer to the average node degree (characteristic of random graph networks). The farther is the ratio from 1, the larger the variations in the node degree (characteristic of scale-free networks).

## 3 Correlation Analysis: Centrality Versus Maximal Clique Size

This section presents the results of correlation coefficient analysis conducted between the centrality values observed for the vertices vis-a-vis the maximal size clique that each vertex is part of. The analysis has been conducted on the six real-world network graphs with respect to centrality and the maximal clique size measured for the vertices in these graphs. The algorithms implemented include those to determine each of the four centrality metrics (Degree, Eigenvector, Betweenness and Closeness) and the exact algorithm to determine the maximal clique size for each vertex in a graph.

Table 1 presents results of the correlation coefficient analysis of the four centrality metrics and the maximal clique size observed for the vertices in each of the six real-world network graphs studied in this paper. Values of correlation coefficient greater than or equal to 0.8 (high correlation) have been indicated in bold; values below 0.5 (low correlation) are indicated in italics; and values between 0.5 and 0.8 (moderate correlation) are indicated in roman. If $\overline{X}$ and $\overline{Y}$ are the average values of the two metrics (say X and Y) observed for the vertices (IDs 1 to $n$, where $n$ is the number of vertices) in the network, the formula used to compute the Correlation Coefficient between two metrics X and Y is as follows:

$$CorrCoeff(X,Y) = \frac{\sum_{ID=1}^{n} (X[ID] - \overline{X}) * (Y[ID] - \overline{Y)}}{\sqrt{\sum_{ID=1}^{N} (X[ID] - \overline{X})^2}\sqrt{\sum_{ID=1}^{N} (Y[ID] - \overline{Y})^2}} \tag{1}$$

**Table 1** Correlation coefficients: centrality metrics and maximal clique size for the nodes

| Network index | Network name (increasing order of spectral radius ratio) | Degree versus clique | Eigenvector versus clique | Closeness versus clique | Betweenness versus clique |
|---|---|---|---|---|---|
| (v) | American College Football Network | *0.32* | *0.35* | *−0.03* | *−0.17* |
| (ii) | Dolphins' Social Network | 0.78 | 0.56 | *0.42* | *0.28* |
| (iii) | US Politics Books Network | 0.70 | 0.75 | *0.32* | *0.37* |
| (i) | Zachary's Karate Club Network | 0.64 | 0.77 | 0.62 | *0.46* |
| (iv) | Word Adjacencies Network | 0.71 | **0.82** | **0.84** | *0.48* |
| (vi) | US Airports 1997 Network | **0.87** | **0.95** | **0.84** | *0.40* |

As one can see in Table 1, in general, the correlation between the centrality metrics and the maximal clique size increases as the spectral radius ratio for node degree increases. This implies, the more scale-free a real-world network is, the higher the correlation between the centrality value and the maximal clique size observed for a node. With several of the real-world networks being mostly scale-free, one could expect these networks to exhibit a similar correlation to that observed in this paper.

The degree-based centrality metrics (degree centrality and eigenvector centrality) have been observed to be very positively and highly correlated with the maximal clique size observed for the nodes. Between the two degree-based centrality metrics, the eigenvector centrality metric shows higher positive correlations to the maximal clique size. This could be attributed to the eigenvector centrality of a node being a measure of both the degree of the node as well as the degrees of its neighbors. That is, a high degree node located in a neighborhood of high degree vertices is more likely to be part of a maximal clique of larger size. In addition, as the networks get increasingly scale-free, nodes with high degree are more likely connected to other similar nodes with high degree (to facilitate an average path length that is almost independent of network size: characteristic of scale-free networks [1] contributing to a positive correlation between degree-based centrality metrics and maximal clique size.

With respect to the two shortest-path based centrality metrics, the betweenness centrality metric is observed to exhibit a low correlation with maximal clique size for all the six real-world network graphs; the correlation coefficient increases as the network becomes increasingly scale-free. In networks with minimal variation in node degree (like the American College Football network that is more closer to a random network), nodes that facilitate shortest-path communication between several node pairs in the network are not part of a larger size clique; on the other hand, nodes that are part of larger size cliques in such random networks exhibit a relatively lower betweenness centrality. Since the degrees of the vertices in random networks are quite comparable to the average node degree, there is no clear ranking of the vertices based on the degree-based centrality metrics and maximal size cliques that they are part of. Also, if at all a vertex ends up being in a larger sized clique in random network graphs, it is more likely not to facilitate shortest path communication between the majority of the vertices (contributing to a negative/zero correlation or at best a low correlation with betweenness centrality). As the network becomes increasingly scale-free, the hubs that facilitate shortest-path communication between any two nodes in the network exhibit higher betweenness and closeness centralities as well as form a clique with other high-degree hubs—exhibiting the ultra small-world property (the average path length is ln(ln $N$), where $N$ is the number of nodes in the network) [1]. The correlation of the closeness centrality values and the maximal clique size values observed for the vertices in real-world network graphs is significantly higher (i.e., positive correlation) for networks that are increasingly scale-free.

Overall, the degree-based centrality metrics exhibit a relatively better correlation with the maximal clique size compared to that of the shortest-path based centrality metrics (especially in networks with low-moderate variation in node degree). For real-world networks that exhibit moderate-high variation in node degree, the shortest-path based centrality metrics (especially closeness centrality) fast catch up with that of the degree-based centrality metrics and exhibit higher levels of positive correlation with maximal clique size. As the networks become increasingly scale-free, the hubs (that facilitate shortest-path communication between any two nodes) are more likely to form the maximum clique for the entire network graph—contributing to higher levels of positive correlation between node centrality and maximal clique size.

## 4 Conclusions

The correlation coefficient analysis studies between the centrality metrics and the maximal clique size for the vertices in the real-world network graphs unravel several significant findings that have been so far not reported in the literature: (i) the degree-based centrality metrics (especially the eigenvector centrality) exhibit a significantly high positive correlation to the maximal clique size as the networks get increasingly scale-free; (ii) the betweenness centrality of the vertices exhibits a low correlation with that of the maximal size cliques the vertices can be part of; (iii) in real-world networks that are close to random network graphs, the centrality metrics exhibit a low correlation to maximal clique size (especially in the case of shortest-path based closeness and betweenness centrality metrics); (iv) for all the four centrality metrics, the extent of positive correlation with maximal clique size increases as the real-world networks become increasingly scale-free.

With the problem of determining maximal clique sizes for individual vertices being computationally time consuming, the approach taken in this paper to study the correlation between maximal clique sizes and centrality can be the first step in identifying positive correlation between cliques/clique size in real-world network graphs to one or more network metrics (like centrality) that can be quickly determined and thereby appropriate inferences can be made about the maximal size cliques of the individual vertices. The degree-based centrality metrics (especially the eigenvector centrality) have been observed to show promising positive correlations to that of maximal clique sizes of the individual vertices, especially as the networks get increasingly scale-free; this observation could form the basis of future research for centrality-clique analysis for complex real-world networks.

## References

1. Newman, M.: Networks: An Introduction, 1st edn. Oxford University Press, Oxford (2010)
2. Strang, G.: Linear Algebra and its Applications, 1st edn. Cengage Learning, Boston (2005)
3. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. MIT Press, Cambridge (2009)
4. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**, 163–177 (2001)
5. Pattabiraman, B., Patwary, M.A., Gebremedhin, A.H., Liao, W.-K., Choudhary, A.: Fast algorithms for the maximum clique problem on massive sparse graphs. In: Bonato, A., Mitzenmacher, M., Pralat, P. (eds.): 10th International Workshop on Algorithms and Models for the Web Graph. Lecture Notes in Computer Science, vol. 8305, pp. 156–169. Springer-Verlag, Berlin Heidelberg New York (2013)
6. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
7. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814–818 (2005)
8. Sadi, S., Oguducu, S., Uyar, A.S.: An efficient community detection method using parallel clique-finding ants. In: Proceedings of IEEE Congress on Evolutionary Computation, pp. 1–7. IEEE, Piscataway NJ (2010)
9. Tomita, E., Kameda, T.: An efficient branch-and-bound algorithm for finding a maximum clique with computational experiments. J. Global Optim. **37**, 95–11 (2007)