International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 5 Issue 7, July-2016

A Centrality Measure for Finding Community in Social Network

Ali Khosrozadeh[1], Mohsen Jahanshahi[2], Shahram Khosrozadeh Ghomi[3]

[1] Department of Computer Engineering, Ayatollah Amoli Branch, Islamic Azad University, Amol, Iran
[2] Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran
[3] Master of Mathematics, Education Office, Babol, Iran

ABSTRACT

This paper presents a centrality measurement and analysis of the social networks for finding online community. The finding of single community in social networks is commonly done using some of the centrality measures employed in social network community finding. The ability that centrality measures have to determine the relative position of a node within a network has been used in previous research work to find communities in social networks using betweenness, closeness and degree centrality measures. It introduces a new metric C-path centrality, and a randomized algorithm for estimating it, and shows empirically that nodes with high C-path centrality have high node betweenness centrality.

Index Terms: Social Network Analysis, Centrality, Communities.

1. INTRODUCTION

Social network analysis [1] views social relationships in terms of network theory consisting of nodes and ties (also called edges, links, or connections). Nodes are the individual Communities [3,7,10] within the networks, and ties are the relationships between the Communities. Measures of centrality [2,8,15] reflect the prominence of communities/units within a network. Within graph theory and network analysis, there are various measures of the centrality of a vertex within a graph that determine the relative importance of a vertex within the graph. The Internet has spawned different types of information sharing systems, including the Web. Recently, online social networks have gained significant popularity and are now among the most popular sites on the Web. Unlike the Web, which is largely organized around content, online social networks [4,5,11] are organized around users. Participating users join a network, publish their profile and (optionally) any content, and create links to any other users with whom they associate. The resulting social network provides a basis for maintaining social relationships, for finding users [9,21] with similar interests, and for locating content and knowledge that has been contributed or endorsed by other users.

Network centrality (or centrality) [8] is used to identify the most important/active people at the center of a network or those that are well connected. Numerous centrality measures such as degree, closeness, betweenness [2,12], information, eigenvector [17], and dependence centrality have been used for characterizing the social behavior and connectedness of nodes within networks. The logic of using centrality measures is that people who are actively involved in one or more communities [16,19] will generally score higher with respect to centrality scores for the corresponding network. Numerous studies in SNA [6] have proposed a diversity of measures to study the communication patterns and the structure of a social network. One of the most studied measures is centrality. Centrality describes a community's relative position within the context of his or her social network [4,13].

2. BACKGROUND AND RELATED WORK

There are four measures of centrality that are widely used in network analysis: degree centrality, betweenness, closeness, and eigenvector centrality.

2.1 Degree Centrality

Degree centrality [8] is defined as the number of links incident upon a node (i.e., the number of ties that a node has). Degree is often interpreted in terms of the immediate risk of node for catching whatever is flowing through the network (such as a virus, or some information). If the network is directed (meaning that ties have direction), then we usually define two separate measures of degree centrality, namely in degree and out degree. In degree is a count of the number of ties directed to the node, and out degree is the number of ties that the node directs to others.

For positive relations such as friendship or advice, we normally interpret in degree as a form of popularity, and out degree as gregariousness. For a graph $G(V, E)$ with n vertices, the degree centrality $C_D(v)$ for vertex v is:

$$C_D(v) = \frac{\deg(v)}{n-1} \qquad (1)$$

The definition of centrality can be extended to graphs. Let v* be the node with highest degree centrality in G. Let $X(Y, Z)$ be the n node connected graph that maximizes the following quantity (with y* being the node with highest degree centrality in X):

$$H = \sum_{j=1}^{|Y|} C_D(y^*) - C_D(y_j) \qquad (2)$$

Then the degree centrality of the graph G is defined as follows:

$$C_D(G) = \sum_{i=1}^{|Y|} (C_D(v^*) - C_D(v_i))/H \qquad (3)$$

H is maximized when the graph X contains one node that is connected to all other nodes and all other nodes are connected only to this one central node (a star graph). In this case

$$H = (n-1)(n-2) \qquad (4)$$

So the degree centrality of G reduces to:

$$C_D(G) = \sum_{i=1}^{|Y|} \frac{C_D(v^*) - C_D(v_i)}{(n-1)(n-2)} \qquad (5)$$

2.2 Betweenness Centrality

Betweenness [8,12] is a centrality measure of a vertex within a graph (there is also edge betweenness, which is not discussed here). Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not.

For a graph $G(V, E)$ with n vertices, the betweenness $C_B(v)$ for vertex v is computed as follows:

1. For each pair of vertices (s, t), compute all shortest paths between them.
2. For each pair of vertices (s, t), determine the fraction of shortest paths that pass through the vertex in question (here, vertex v).
3. Sum this fraction over all pairs of vertices (s, t).
Or, more succinctly: [2]

$$C_B(v) = \sum_{s \neq v \neq t \in V}^{|Y|} \sigma_{st}(v)/\sigma_{st} \qquad (7)$$

where σ_{st} is the number of shortest paths from s to t, and $\sigma_{st}(v)$ is the number of shortest paths from s to t that pass through a vertex v.

Calculating the betweenness and closeness centralities of all the vertices in a graph involves calculating the shortest paths between all pairs of vertices on a graph. In calculating betweenness and closeness centralities of all vertices in a graph, it is assumed that graphs are undirected and connected with the allowance of loops and multiple edges. When specifically dealing with network graphs, oftentimes graphs are without loops or multiple edges to maintain simple relationships (where edges represent connections between two people or vertices). In this case, using Brandi's algorithm [22] will divide final centrality scores by 2 to account for each shortest path being counted twice.

2.3 Closeness Centrality

In topology and related areas in mathematics, closeness [8] is one of the basic concepts in a topological space. Intuitively we say two sets are close if they are arbitrarily near to each other. The concept can be defined naturally in a metric space where a notion of distance between elements of the space is defined, but it can be generalized to

topological spaces where we have no concrete way to measure distances. In graph theory closeness is a centrality measure of a vertex within a graph. Vertices that are "shallow" to other vertices (that is, those that tend to have short geodesic distances to other vertices within the graph) have higher closeness. Closeness is preferred in network analysis to mean shortest-path length, as it gives higher values to more central vertices, and so is usually positively associated with other measures such as degree.

In the network theory, closeness is a sophisticated measure of centrality. It is defined as the mean geodesic distance (i.e., the shortest path) between a vertex v and all other vertices reachable from it:

$$\sum_{t \in V \setminus v} \frac{d_G(v,t)}{n-1} \qquad (8)$$

Where $n \geq 2$ is the size of the network's "connectivity component" V reachable from v. Closeness can be regarded as a measure of how long it will take information to spread from a given vertex to other reachable vertices in the network. Some define closeness to be the reciprocal of this quantity, but either way the information communicated is the same (this time estimating the speed instead of the time span). The closeness $C_c(v)$ for a vertex v is the reciprocal of the sum of geodesic distances to all other vertices of V:

$$C_c(v) = \frac{1}{\sum_{t \in V \setminus v} d_G(v,t)} \qquad (9)$$

Different methods and algorithms can be introduced to measure closeness, like the random-walk centrality [14] introduced by Noh and Rieger (2003) that is a measure of the speed with which randomly walking messages reach a vertex from elsewhere in the network - a sort of random-walk version of closeness centrality. The information centrality [8,18] of Stephenson and Zelen (1989) is another closeness measure, which bears some similarity to that of Noh and Rieger. In essence it measures the harmonic mean length of paths ending at a vertex i, which is smaller if i has many short paths connecting it to other vertices. Dangalchev (2006), in order to measure the network vulnerability, modifies the definition for closeness so it can be used for disconnected graphs and the total closeness is easier to calculate:

$$C_c(v) = \sum_{t \in V \setminus v} 2^{-d_G(v,t)} \qquad (10)$$

An extension to networks with disconnected components has been proposed by Opsahl (2010).

2.4 Eigenvector Centrality

Eigenvector centrality [17] is a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Betweenness centrality is mostly used to find and measure subgroup and community membership whereas degree and closeness centrality are used for characterizing influential members. Although network centrality measures are easy to calculate using computer programs such as Pajek [25] and UCINET [24], there has been no consensus among researchers as to the most meaningful centrality measure to use for finding subgroup members. In extremely large social networks, computational efficiency may become an issue in selecting which centrality measure to use. With respect to three commonly used centrality measures, degree centrality is the easiest to calculate, closeness centrality is more complex and betweenness centrality has the highest calculation complexity.

3. PROPOSED MEASUREMENT METHODOLOGY

Assume the traversal of a message (e.g., news or rumour) originating from some source s over a network and intending to finally reach some destination t in the network along a path, and assume that each node in the network has only its own local view (i.e., has information only of its outgoing neighbors). Thus, when the message is at a current node v, the node v forwards the message based on its local view to one of its outgoing neighbors chosen uniformly at random. The message continues to travel in this manner until it reaches the destination node t, and then stops.

The notion of C-path centrality is based on a similar assumption regarding the random traversal of a message from a source s. However, we make two further assumptions in order to reduce the computation time without deviating much from the above random walk model. First, we consider message traversals along simple paths only, i.e., paths in which vertices do not repeat. As non-simple paths do not correspond to the intuitive notion of ideal message traversals in a

social network, their consideration in the computation of centrality indices is a noisy factor. To discount non-simple paths, we assume that each intermediate node v on a partially traversed path forwards the message to a neighbor chosen randomly, with probability inversely proportional to edge weights, from the current set of unvisited neighbors; the message traversal is assumed to stop if all the outgoing neighbors of the current node v already appear in the path up to v. Although choosing a random neighbor in this manner at each step requires the premise that the message carries the history of the path traversed so far, this premise is needed to express the average contribution of any simple path in the overall information flow and to efficiently simulate such random simple paths. Second, we assume that the message traversals are only along paths of at most C links (edges), where C is a parameter dependent on the network. It has been found in many studies on social networks that message traversals typically take paths containing few links [23], and so this seems to be a reasonable assumption in the context of social networks. Based on these assumptions, we define C-path centrality:

DEFINITION (C-Path Centrality) for every vertex v of a graph $G(V, E)$, the C-path centrality $C_c(v)$ of v is defined as the sum, over all possible source nodes s, of the probability that a message originating from s goes through v, assuming that the message traversals are only along random simple paths of at most C edges.

3.1 Estimating C-Path Centrality

We present a randomized approximation algorithm for estimating the C-path centrality of all vertices in any graph. The algorithm takes as input a graph $G(V, E)$, a nonnegative weight function W on the edges of G, and parameters $\alpha \in [-1/2,1/2]$ and integer $C = f(m, n)$, and runs in time $O(K^3 n^{2-2\alpha} \ln n)$. For each vertex v, it outputs an estimate of $C_c(v)$ up to an additive error of $n^{\frac{1}{2+\alpha}}$ with probability at least $1 - \frac{1}{n^2}$. We refer to this algorithm as Randomized-Approximate C path.
Input: Graph $G(V, E)$, Array W of edge weights, $\alpha \in [-1/2,1/2]$ and integer C.
Output: Array K of C-path centrality estimates.

```
begin
for each v ∈ V do
    count[v] ← 0;
    Explored[v] ← false;
end
/* S is a stack and n = |V|*/
T ← 2C²n^(1−2α) ln n;
S ← ∅;
for i ← 1 to T do
     /* Simulate a message traversal from s containing e edges */
    s ← a vertex chosen uniformly at random from V;
    e ← an integer chosen uniformly at random from [1,C];
    Explored[s] ← true;
    push s to S;
    j ← 1;
    while (j ≤ e and ∃(s; u) ∈ E such that !Explored[u]) do
          v ← a vertex chosen randomly from { u|(s, u) ∈ E  and
          !Explored[u] } with probability proportional to 1/W(s; v);
        Explored[v] ← true;
        push v to S;
        count[v] ← count[v] + 1;
        s ← v;
        j ← j + 1;
    end
    /* Reinitialize Explored[v] to false */
    while S is nonempty do
        pop v ← S;
        Explored[v] ← false;
    end
end
for each v ∈ V do
    K[v] ← C_n.count[v]/T ;
end
return K;
end
```

The algorithm performs $T = 2C^2 n^{1-2\alpha} \ln n$ iterations (the expression for T comes from the analysis of the algorithm). In each iteration, a start vertex $s \in V$ and a walk length $e \in [1, C]$ are chosen uniformly at random, and then a random walk consisting of e edges from s is performed that essentially simulates a message traversal from s in G using the assumption made in Definition. The number of times any vertex v is visited over all the random walks is recorded in a variable count[v]. The estimated C-path centrality K[v] of any vertex v is then defined as the scaled average of the times v is visited over T walks:

$$K[v] \leftarrow C_n . count[v]/T$$

3.2 Example

Above mentioned centrality measures work on various nodes I_1,I_3,S_1,S_4,W_1,W_2,W_3,W_4,W_5,W_6,W_7,W_8,W_9 and find community using influential nodes in following manner:

I_1 - W_1 - W_2 - W_3 - W_4
I_3
W_1 - I_1 - W_2 - W_3 - W_4 - W_5 - S_1
W_2 - I_1 - W_1 - W_3 - W_4 - S_1
W_3 - I_1 - W_1 - W_2 - W_4 - W_5 - S_1
W_4 - I_1 - W_1 - W_2 - W_3 - W_5 - S_1
W_5 - W_1 - W_3 - W_4 - W_7 - S_1
W_6 - W_7 - W_8 - W_9
W_7 - W_5 - W_6 - W_8 - W_9 - S_4
W_8 - W_6 - W_7 - W_9 - S_4
W_9 - W_6 - W_7 - W_8 - S_4
S_1 - W_1 - W_2 - W_3 - W_4 - W_5
S_2
S_4 - W_7 - W_8 - W_9

Table1: Comparison of various centrality measures using UCINET Simulator

Node	Degree	Closeness	Betweenness	Eigenvector	C-path
I_1	30.769	23.636	0.000	43.368	4.00
I_3	0.000	0.000	-----	0.000	0.00
S_1	38.462	26.531	1.923	52.043	6.75
S_2	0.000	-----	0.000	0.000	0.00
S_4	23.077	23.636	0.000	4.070	3.00
W_1	46.154	27.083	4.808	58.960	10.583
W_2	38.462	24.074	0.321	51.669	5.05
W_3	46.154	27.083	4.808	58.960	10.583
W_4	46.154	27.083	4.808	58.960	10.583
W_5	38.462	28.889	38.462	45.718	31.000
W_6	23.077	23.636	0.000	4.070	3.000
W_7	38.462	27.660	36.325	12.011	37.667
W_8	30.769	24.074	0.427	4.719	4.667
W_9	30.769	24.074	0.427	4.719	4.667

CONCLUSION

This paper present a new approach for identifying highly influential nodes based on their C-path centrality score, according to the following observations. First, we observe that the value of the C-path centrality is irrelevant: it is the relative importance of communities (as measured by C-path centrality) that matters. Second, we observe that for the vast majority of applications, it is sufficient to identify categories of nodes of similar importance. Third, we observe that distant communities in social networks are unlikely to influence each other.

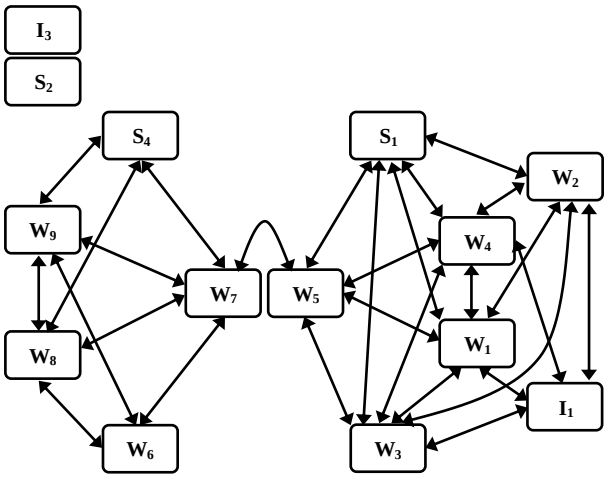

Fig1: Visualization of centrality measures work on various communities using UCINET simulator

REFERENCES

[1]. Garton L, Haythornthwaite C, Wellman B. "Studying Online Social Networks". J Comput Mediated Commun, 3(1):1–30, 1997.
[2]. L. Freeman. "A Set of Measures of Centrality Based on Betweenness". Sociometry, 40(1):35–41, 1977.
[3]. Estrada E, Rodriguez-Velazquez AJ. "Subgraph Centrality in Complex Networks". Phys Rev E 71:056103, 2005.
[4]. Backstrom L. "Group Formation in Large Social Networks: Membership, Growth, and Evolution". In: KDD 06: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, ACM Press, pp 44–54, 2006.
[5]. Burt R. "Toward a Structural Theory of Action: Network Models of Social Structure, Perception and Action". Academic, New York, 1982.
[6]. Carrington PJ, Scott J, Wasserman S. "Models and Methods in Social Network Analysis". Cambridge University Press, New York, NY, USA, 2006.
[7]. Danon L, Duch J, Diaz-Guilera A, Arenas A. "Comparing Community Structure Identification". J Stat Mech Theor Exp: P09008, 2005.
[8]. Freeman CL. "Centrality in Social Networks: Conceptual Clarification". Social Networks 1:215–239, 1978.
[9]. Clauset A. "Finding Local Community Structure in Networks. Phys Rev E 72:026132, 2005.
[10]. Du N, Wu B, Pei X, Wang B, Xu L. "Community Detection in Large-Scale Social Networks". In WebKDD/SNA-KDD'07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, New York, NY, USA, pp 16–25, 2007.
[11]. N. Friedkin. "Horizons of Observability and Limits of Informal Control in Organizations". Social Forces, 62(1):57–77, 1983.
[12]. G. Kahng, E. Oh, B. Kahng, D. Kim. "Betweenness Centrality Correlation in Social Networks". Phys Rev E, 67:01710–1, 2003.
[13]. Danon L, Duch J, Diaz-Guilera A, Arenas A. "Comparing Community Structure Identification". J Stat Mech Theor Exp: P09008, 2005.
[14]. M. Newman. "A Measure of Betweenness Centrality Based on Random Walks". Social Networks, 27(1):39–54, 2005.
[15]. K. Stephenson, M. Zelen. "Rethinking Centrality: Methods and Examples". Social Networks, 11:1–37, 1989.
[16]. Newman EJM, Girvan M. "Finding and Evaluating Community Structure in Networks". Phys Rev E 69:026113, 2004.
[17]. Ruhnau B. "Eigenvector Centrality, a Node Centrality?". Social Networks, 22(4):357–365, 2000.
[18]. Fortunato S, Latora V, Marchiori M. "Method to Find Community Structures Based on Information Centrality". Phys Rev E (Stat Nonlinear, Soft Matter Phys) 70(5):056104, 2004.
[19]. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. "Defining and Identifying Communities in Networks". Proc Natl Acad Sci USA 101(9):2658– 2663, 2004.
[20]. Tantipathananandh C, Berger-Wolf YT, Kempe D. "A Framework for Community Identification in Dynamic Social Networks". In: KDD'07: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, pp 717–726, 2007.
[21]. Traud LA, Kelsic DE, Mucha JP, Porter AM. "Community Structure in Online Collegiate Social Networks". American Physical Society, 2009 APS March Meeting, March 16–20, 2009.
[22]. U. Brandes. "A Faster Algorithm for Betweenness Centrality". Journal of Mathematical Sociology, 25(2):163–177, 2001.
[23]. U. Brandes, C. Pich. "Centrality Estimation in Large Networks". I. J. of Bifurcation and Chaos, 17(7):2303–2318, 2007.
[24]. Borgatti SP, Everett GM, Freeman CL. "Ucinet for Windows: Software for Social Network Analysis". Analytic Technologies, Harvard, USA Science BV, Amsterdam, the Netherlands, pp 107–117, 2002.
[25]. De Nooy W, Mrvar A, Batagelj V. "Exploratory Social Network Analysis with Pajek". Cambridge University Press, New York, USA, 2005.