An Efficient Community Detection Method Based on Path Proximity and Local Edge Betweenness Centrality

Ye Zhu, Jun Gong, Simeng Wu

School of Software, Jiangxi Normal University, Jiangxi Province, China

ABSTRACT. *Centrality is a measure index evaluating node and edge importance, and edge betweenness centrality is the most important among those. The length of shortest path between two nodes shows their relationship. When the path is longer, their relationship is weaker. Based on this idea, this paper think that two nodes have no relationship beyond some range. This paper introduces local edge betweenness centrality within K steps and weight of path length, and puts forward a new centrality evaluation method. Based on this, our method utilizes the core idea of GN algorithm, and puts forward a new community detection algorithm called Dist-K. Its performance and result are compared with other existing methods, as applied to different well-known instances of complex networks. Our method has better modularity, normalized mutual information, adjusted rand index and accuracy, which are widely used to measure community structures.*

KEYWORDS: *local edge betweenness centrality, community detection, path length*

1. Introduction

The new era is an era of coexistence of complexity and network, that is, the complex network era. In the whole ecological environment, biological individuals are closely related to their own environment and interact with each other. Individuals are regarded as nodes, and their interactions are regarded as links, namely, they jointly constitute the network structure. The connections in the network are crisscrossed and complex, which is the most basic characteristic of the network. In addition, complex networks also have three characteristics: small-world characteristics, scale-free characteristics and superfamily characteristics.

The field of complex network is one of the hot spots of current research. In this paper, we focus on the computation of improved edge betweenness centrality in undirected unweighted static networks, then combine the core idea of GN algorithm to get the result of community detection. The following three aspects are introduced respectively: network mechanism model, betweenness centrality and community detection algorithm.

So far, network models that are relatively complete and formed include regular network, random network [1], small world network [2] and scale-free network [3]. Centrality reflects the relative importance of each node in complex network, and in complex network analysis, the characterization methods of centrality mainly includes degree centrality, betweenness centrality, closeness centrality and eigenvector centrality [4]. At present, there are many bases for dividing community detection algorithms. They are divided into static and dynamic algorithms according to whether they change over time, and overlapping and non-overlapping algorithms according to whether they can detect overlapping communities. In this paper, static non-overlapping community detection algorithms are mainly involved.

2. The Quantifying and Identifying Methods

A. Quantifying the community structure

The modularity was used to measure the goodness of a partition of network, which was proposed by Newman and Girvan [5] in 2004, and the modularity can be formalized as:

$$Q = \sum_i (e_{ii} - a_i^2) \tag{1}$$

The range of modularity is [-0.5,1), the larger the modularity is, the better the result of community detection will be. In the actual network, the modularity is usually between 0.3 and 0.7, and the probability of greater than 0.7 is very small.

The normalized mutual information (NMI) was used to measure the goodness of a partition of network, which was proposed by Vinh N X, Epps J and Bailey J [6] in 2010, and the normalized mutual information can be formalized as:

$$NMI(X;Y) = 2\frac{I(X;Y)}{H(X)+H(Y)} \tag{2}$$

$$I(X;Y) = \sum_x \sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{3}$$

$$H(X) = \sum_i h(P(X = X_i)) \tag{4}$$

$$h(x) = -x * \log(x) \tag{5}$$

The value range of NMI is [0,1], where the joint distribution of two random variables $(X;Y)$ is $p(x,y)$, and the edge distribution are $p(x,y)$, $p(y)$, the larger the NMI is, the better the result of community detection will be.

The adjusted rand index (ARI) was used to measure the goodness of a partition of network, which was proposed by Danon L and Duch J [7] in 2005, and the adjusted rand index can be formalized as:

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)} \tag{6}$$

$$RI = \frac{a+b}{C_m^2} \tag{7}$$

Where parameter a represents the number of nodes where the real partition result and the predicted partition result are both wrong, parameter b represents the number of nodes where the real partition result and the predicted partition result are both right, and parameter m represents the number of nodes. the larger the ARI is, the better the result of community detection will be.

B. Identifying the community structure

GN community detection algorithm regards all nodes in the network as a community, calculates the edge betweenness centrality of all edges in the network, removes the edge with the largest edge betweenness centrality from the network and records the connected component at this time. If the number of connected components is more than the number of connected components recorded last time, that is, a new connected component is created, the modularity is calculated and recorded at this time. Each time an edge is removed, the edge betweenness centrality of the remaining edges need to be recalculated until all edges are removed. Find the connected component with the maximum modularity, which is the result of community discovery algorithm. The core of the algorithm is the computation of edge betweenness centrality, while the traditional computation of edge betweenness centrality is the accumulative computation of edge betweenness centrality along the shortest paths. On the one hand, the information will spread along the shortest paths, and it will gradually decline with the propagation, so the possibility of the information to a distant location is very small. Therefore, when calculating the edge betweenness centrality, it may introduce error to include the distant of pair of nodes into the calculation. On the other hand, with a certain amount of information, each edge accumulates different amount of information under different path lengths. From these two aspects, the interference of the long path and different path lengths to the measurement of the edge betweenness centrality is introduced.

3. Implementation and Results

The experiment of this paper uses Karate network [8], Dolphins network [9-10], Polbooks network [11] and Football network [12]. The basic information of the data sets are shown in Table 1.

Table 1 The basic information of the data sets

Data set	Number of Nodes	Number of Edges	Diameter of Network	Number of Communities
Karate	34	78	5	2
Dolphin	62	159	8	2
Polbooks	105	441	7	3
Football	115	613	4	12

Figure 1 shows the cumulative probability of distance between the pair of nodes in real network, it is obvious that Karate network can reach 90% nodes in the network within 3 steps, Dolphins and Polbooks network can reach 90% nodes in the network within 4 or 5 steps, and When K is 3, Football network calculates an approximate global betweenness centrality, which is not consistent with the purpose of experiment, so when K is 2, the result of experiment is more reasonable.

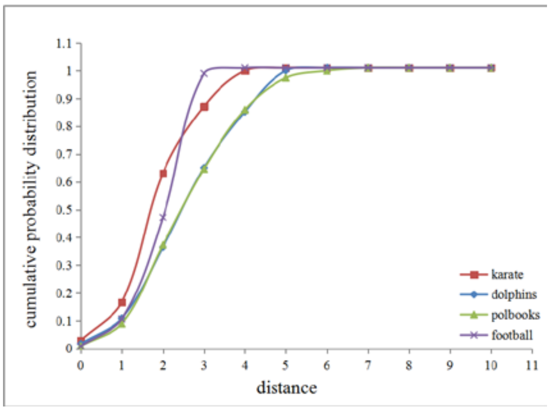

Figure. 1 The cumulative probability distribution of distance between the pair of nodes in Karate network, Dolphins network, Polbooks network and Football network

Figure 2 shows the change of modularity under different K values in real network, when K exceeds the longest distance between the pair of nodes, the global edge betweenness centrality is calculated, so K has to less than the diameter of network.Considering efficiency, smaller K requires less computation.When K reaches 3, the modularity of Karate network is the largest and unchanged, so in the best case, K is 3. When K reaches 5, the modularity of Dolphins network is the largest and unchanged, so in the best case, K is 5. Polbooks network has the largest modularity when K is 4, so it is the best case. When K reaches 2, the modularity of Football network is the largest and unchanged, so in the best case, K is 2.

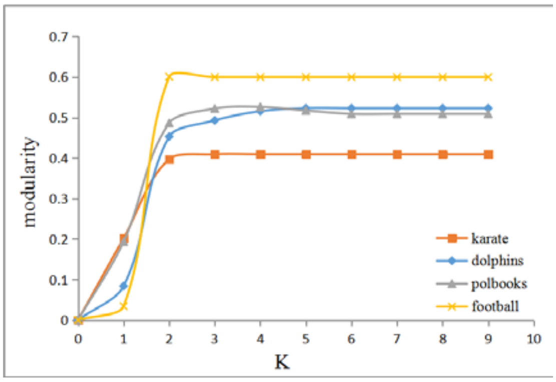

Figure. 2 The change of modularity under different K values in Karate network, Dolphins network, Polbooks network and Football network

The results of community detection of Karate network, Dolphins network, Polbooks network and Football network are shown in Figure 3-6. The results are obtained by applying different K in different networks to the community detection algorithm in this paper.

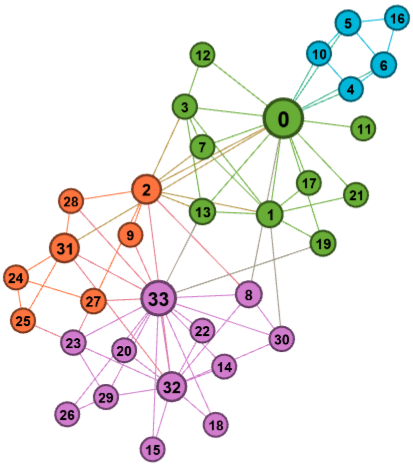

Figure. 3 The community detection result of Karate

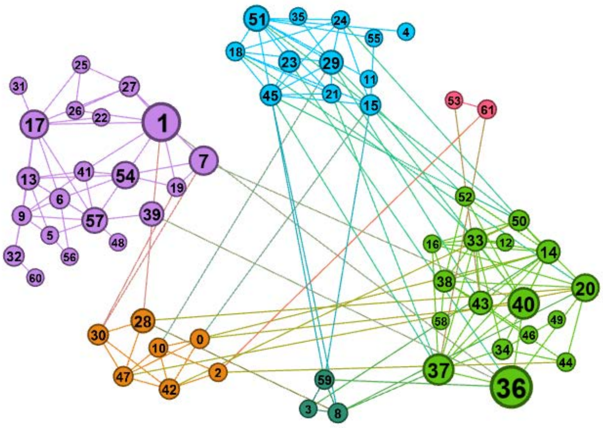

Figure. 4 The community detection result of Dolphins

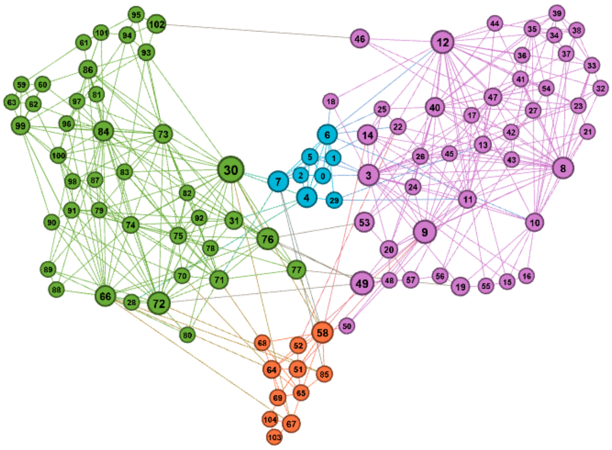

Figure. 5 The community detection result of Polbooks

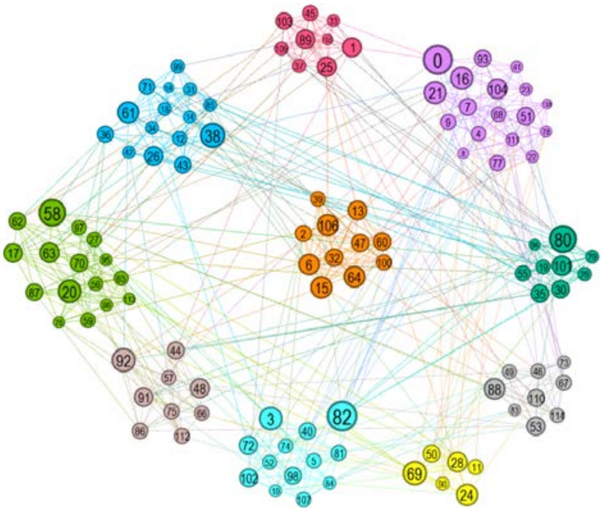

Figure. 6 The community detection result of Football

Table 2 Results compared with classical community detection algorithms

Data set	Index	GN	FN	Louvain	LPA	Walktrap	Dist_K
Karate	C_N	5	3	4	1-3	5	4
	Q	0.4013	0.3807	0.4188	0.3147	0.3532	0.40935
	ARI	0.3916	0.5684	0.3922	-	0.3207	0.40644
	NMI	0.4851	0.5646	0.4900	-	0.4899	0.5005
	Accuracy	0.4412	0.2647	0.3235	-	0.3824	0.4706
Dolphins	C_N	5	4	5	2-5	4	6
	Q	0.5194	0.4955	0.5233	0.4920	0.4888	0.5230
	ARI	0.3430	0.3983	0.2941	-	0.3706	0.2973
	NMI	0.5050	0.5260	0.4468	-	0.4984	0.4721
	Accuracy	0.4194	0.6774	0.3871	-	0.6452	0.4194
Polbooks	C_N	5	4	4	1-4	4	4
	Q	0.5168	0.5020	0.5204	0.3801	0.5070	0.5262
	ARI	0.6823	0.6379	0.5580	-	0.6534	0.6649
	NMI	0.5584	0.5308	0.5121	-	0.5427	0.5537
	Accuracy	0.8095	0.0667	0.7238	-	0.7905	0.8
Football	C_N	10	6	10	8-13	10	12
	Q	0.5996	0.5497	0.6046	0.5819	0.6029	0.6005
	ARI	0.7781	0.4741	0.8069	-	0.8154	0.8967
	NMI	0.8789	0.6977	0.8903	-	0.8874	0.9242
	Accuracy	0.2174	0.1826	0.2174	-	0.2087	0.3913

Compared with other classical community detection algorithms, modularity(Q), normalized mutual information (NMI), adjusted rand index (ARI) and Accuracy are used for evaluation of the quality of community detection result. The higher the Q, NMI, ARI and Accuracy are, the better the results are, which are shown in Table 2. C_N represents the number of communities in the result of community detection algorithms.

Table 2 shows the experimental results of Dist_K algorithm and 5 classical community discovery algorithms in data sets of real networks with labels. Due to the instability of LPA algorithm, only the average of modularity in 30 experiments is given. As you can see, in general, the Dist_K algorithm performs better in these data sets of real networks.

References

[1] Pàl Erds, A. R. Rényi. On random graphs I [J]. Publicationes Mathematicae, 1959, 6: 290-297.

[2] Watts D, Strogatz S. Collective dynamics of 'small-world' networks [J]. Nature, 1998, 393 (6684): P.440-442.

[3] Barabasi A L, Albert R. Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509-512 [J]. ence, 1999, 286 (5439): 509-512.

[4] Merrer EL, Tredan G. Centralities: Capturing the fuzzy notion of importance in social graphs. In: Proc. of the European Conf. on Computer Systems. 2009. 33-38 [doi:10.1145/1578002.1578008]

[5] Newman M E. Fast algorithm for detecting community structure in networks[J]. Phys Rev E Stat Nonlin Soft Matter Phys, 2004, 69(6 Pt 2): 066133.

[6] Vinh N X, Epps J, Bailey J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance [M]. JMLR.org, 2010.

[7] Danon L, Duch J, Diaz-Guilera A, et al. Comparing community structure identification [J]. Journal of Statal Mechanics, 2005, 2005 (09): 09008.

[8] Wayne, W, Zachary. An Information Flow Model for Conflict and Fission in Small Groups [J]. Journal of Anthropological Research, 1977.

[9] Lusseau D, Schneider K, Boisseau O J, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations [J]. Behavioral Ecology & Sociobiology, 2003, 54 (4): 396-405.

[10] Lusseau D, Newman M E J. Identifying the role that animals play in their social networks. [J]. Proceedings of the Royal Society B Biological ences, 2004, 271 (Suppl_6): S477.

[11] Shen H W. Community Structure of Complex Networks [J]. Complex Systems and Complexity ence, 2011.

[12] Newman M E J. Modularity and community structure in networks [C] // 2006 APS March Meeting. American Physical Society, 2006.