Differential Betweenness in Complex Networks Clustering

Alberto Ochoa[1] and Leticia Arco[2]

[1] Institute of Cybernetics, Mathematics and Physics
ochoa@icmf.inf.cu
[2] Central University of Las Villas
leticiaa@uclv.edu.cu

Abstract. We propose a novel metric for measuring the degree of edge centrality in complex networks clustering, a task commonly called community detection in the analysis of social, biological and information networks. The metric, which has been called differential betweenness, has some unexpected and interesting properties that might help us to create better clustering algorithms. We compare our measure with the shortest path edge betweenness of Girvan and Newman and found that it can be more accurate and robust without requiring the costly recalculation step the other measure needs.

Keywords: graph clustering, betweenness centrality, complex networks.

1 Introduction

When one uses a graph to model a real problem –entities are represented by nodes and their relationships by edges, finding clusters of well connected nodes is clearly a pattern recognition task of major importance. The interesting point here, is that one immediately discovers that the popular clustering techniques often fail in dealing with the so called complex networks. This is the case of community detection in social, biological and information networks.

In the last eight years, there has been a lot of activity in the above mentioned fields regarding theoretical developments in community detection and its applications. M. J. E. Newman is a prolific author on the topic. Thus, the interested reader is referred to some of his works and the references therein for a rather complete exposition about the methods used so far [1,2,3,4,6].

At the beginning, the research was more or less a kind of feature extraction process intended to capture notions of vertex and edge centrality. In the case relevant to this paper, the measurements are used to decide which edges and when should be removed from the graph to discover its internal structure. These were hierarchical divisive clustering algorithms combined with a validity index necessary to choose the best decomposition. In complex networks research, the most widely used measure for determining the quality of a community partitioning was proposed by Girvan and Newman [5]. It was called modularity index,

J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 227–234, 2008.

Q. Another important contribution of these authors was the so called edge betweenness centrality and the most popular algorithm that used it: the Girvan and Newman betweenness algorithm (GN). The major shortcoming of GN and similar algorithms was their high computational cost.

In the second period of development of community detection, many leading researchers abandoned the initial ideas and switched to the direct optimization of the modularity index (see for example [4]). Although these methods are interesting and promising we are still not convinced that the former approaches were superseded by them. This believe is the first motivation of our work. We think that the forward movement made by the research community was premature.

In our opinion, several important issues regarding the metrics were missed in the early contributions. Better designed metrics, in combination with some optimization and/or machine learning (classification) tools will lead to faster and more accurate detection algorithms. In this paper we take one step in this direction. We present the differential betweenness and show that this novel metric has some unexpected and interesting properties that can help us to create powerful detection algorithms for large complex networks.

The outline of the paper is as follows. In Sect. 2 we present a brief introduction to the problem of communities detection in complex networks. Section 3 describes the need for new betweenness metrics and introduces our proposal: the differential betweenness. Section 4 supports the claims of the previous section by presenting an empirical comparative study of the shortest-path betweenness and our metric in the case of the famous Zachary's karate club network. Section 5, presents a short discussion on the time complexity of the proposed metric. Finally, Sect. 6 presents our conclusions.

2 Community Detection in Complex Networks

For the purposes of this paper we have chosen, as starting point, the classic Girvan Newman community detection algorithm.

The GN detection is based in a metric called shortest path edge betweenness which counts the number of geodesics between pairs of vertexes that run along an edge. The general form of the algorithm is as follows: 1) Calculate betweenness scores for all edges in the network. 2) Find the edge with the highest score and remove it from the network. 3) Recalculate betweenness for all remaining edges. 4) Repeat from step 2. This algorithm is slow, it needs time $O(m^2n)$ for dense and $O(n^3)$ for sparse networks –m edges and n nodes.

The step 3 is critical and adds one order of complexity with respect to m, which means a worst case addition of order n^2. However, it has been considered the heart of the algorithm. In [5] Newman wrote:

> "... our studies indicate that, once one hits on the idea of using betweenness measures to weight edges, the exact measure one uses appears not to influence the results highly. The recalculation step, on the other hand, is absolutely crucial to the operation of our methods. This step

> was missing from previous attempts at solving the clustering problem using divisive algorithms, and yet without it the results are very poor indeed, failing to find known community structure even in the simplest of cases."

One of the new contributions of our work is to present a different perspective on the above issue. A major shift of the importance from the recalculation step to the adequacy of the metric. In section 4, we will show that a different measure can be less sensitive to recalculation.

3 Differential Betweenness

There is a fundamental issue dealing with the difference between the objective notion of betweenness and what can be possibly measured of it. What we understand by an objective notion is the property that some edges have of "mediating" between communities regardless of whether we can adequately measure that capacity or not. The comparison we have made with a measuring process is intentional. This leads us to an interesting reflection on the quality of the measuring device. If we cannot measure what we want, we cannot expect our inferences to be correct. When the edge betweenness is measured, some neighboring edges and others not so close can become noise factors. This effect is what the GN algorithm tries to avoid using the recalculation step. In other words, we could say that there is some complex dependency between the measurements. Not even will the recalculation process be able to guarantee a correct estimation.

It is important to stress that we have been referring to the independence of the measuring process and not to the independence of the actual betweenness values. The ideal thing would be to develop a measuring method that would not depend on the other edges or not even on the network topology. The analysis of the second type of independence having to do with the network topology also leads us to another important reflection: what the betweenness measure should be like. The values of the edge betweenness are independent of the values of the rest of the network edges. It depends on the values of a certain edge neighborhood. This "Markov" property allows measuring to be made locally, with the subsequent benefits in terms of algorithm efficiency.

The previous considerations, and others no included due to space constraints, turn out to be the conceptual and methodological starting point of the new betweenness measure that we now introduce formally.

In this paper, the notation $\|a -_s b\|$ will represent the length of the geodesics between the pair of nodes a and b, whereas $\|a -_s b \,|i - j\,\|$ will refer to the length of the shortest path between a and b that runs along the edge $i - j$. The following definition introduces the locality in our approach.

Definition 1 (c-neighborhood). *A c-neighborhood of the edge $i - j$ in the graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$, is the subgraph $\mathsf{G}_{c,i-j} = (\mathsf{V}_{c,i-j}, \mathsf{E}_{c,i-j})$, where:*

$$\mathsf{V}_{c,i-j} = \{v \in \mathsf{V} \mid \|v -_s i\| \leq c \vee \|v -_s j\| \leq c\}$$
$$\mathsf{E}_{c,i-j} = \{(i,j) \in \mathsf{E} \mid i,j \in \mathsf{V}_{c,i-j}\}$$

We declare that the results of the paper remain valid with many other definitions of c-neighborhood.

Definition 2 (differential geodesic). *Let a, b, i and j be nodes of the graph* $\mathsf{G} = (\mathsf{V}, \mathsf{E})$. *The function ζ is defined as follows:*

$$\zeta\left(i, j \left| a, b\right.\right) = \left\| a -_s b \left| i - j \right. \right\| - \left\| a -_s b \right\| \tag{1}$$

We say that (1) denotes the differential geodesic of the pair of nodes a and b with respect to the edge $i - j$. It is non negative in its domain of definition, and reaches its minimum value when the geodesic between a and b runs along $i - j$. A more general interpretation of the differential geodesic assumes that $i - j$ is not an edge but a shortest path. However, the analysis of this case is beyond the scope of this paper.

The next definition is key for the introduction of the differential betweenness.

Definition 3 (λ-betweenness). *We call λ-betweenness of the edge $i - j$ given the pair of nodes (a, b) the expression*

$$B_\lambda\left(i, j \left| a, b\right.\right) = e^{-\lambda \zeta\left(i, j \left| a, b\right.\right)}$$

where the function $\zeta\left(i, j \left| a, b\right.\right)$ follows definition 2.

Note that $0 < B_\lambda\left(i, j \left| a, b\right.\right) \leq 1$ and that for any finite λ the maximum is attained when the geodesic runs along $i - j$.

Definition 4 (Differential Betweenness). *Let the graph* $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ *be given. The magnitude $DB_{\lambda,c}(i, j)$* given by the following expression

$$DB_{\lambda,c}(i, j) = \sum_{a,b \in \mathsf{V}_{c,i-j}} B_\lambda\left(i, j \left| a, b\right.\right) = \sum_{a,b \in \mathsf{V}_{c,i-j}} e^{-\lambda \zeta\left(i, j \left| a, b\right.\right)},$$

is what we have called differential betweenness of the edge $i - j$ in the c-neighborhood $\mathsf{V}_{c,i-j}$ with parameter λ.

The above definition is rather general and is in fact just one of the possible implementations of the differential betweenness idea. For example, the definition for weighted networks is similar. A deep discussion of all the related issues is beyond the scope of this paper and will be presented elsewhere. We would like to stress once again that the aim of the paper is the presentation of the new measure together with evidences that highlight some of its remarkable properties. This is precisely what we will do in the following section.

4 Differential Betweenness in the Zachary's Network

The aim of this section is to present one simple example where the differential betweenness captures more information about the structure of a network

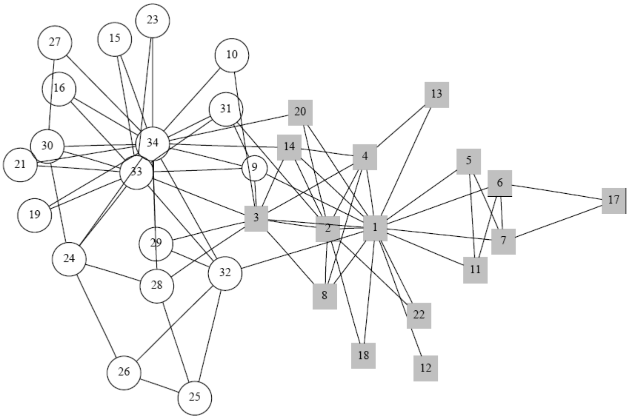

Fig. 1. Zachary's karate club network divided in two communities

than the shortest-path (GN) betweenness. We take a famous benchmark: the Zachary's karate club problem. This network has 34 vertexes, 78 edges and two communities connected by 10 bridges (see Fig. 1). We refer the interested reader to the cited works for more information on this benchmark.

We computed three betweenness values for each edge $i \sim j$: $DB_{0.01,2}(i,j)$, $DB_{0.01,4}(i,j)$ and $GN(i,j)$. Note that in this case, $c = 4$ amounts to use the whole graph as the edge c-neighborhood. Then these arrays were sorted in decreasing order. Besides the betweenness values, Table 1 shows the position, P, of the inter-communities bridges according to the mentioned order.

Table 1. Betweenness values for the bridges of the Zachary's network

Bridge	P	DB_2	P	DB_4	P	GN	*Bridge*	P	DB_2	P	DB_4	P	GN
3-9	2	396.37	13	455.82	5	41.65	3-33	22	276.03	49	290.89	60	8.33
3-29	14	317.03	37	330.83	57	10.47	3-10	23	275.51	42	317.40	66	5.15
3-28	16	313.39	39	327.21	75	2.37	2-31	24	255.78	38	329.76	74	2.54
1-9	18	301.29	44	301.29	76	1.89	20-34	25	253.83	50	273.63	24	19.49
1-32	20	295.02	48	295.02	65	5.5	14-34	28	231.93	52	249.78	35	16.61

There are a number of interesting things to note. First of all, it is remarkable that for the GN betweenness the positions of the bridges rather evenly spread across the interval $[1, 78]$. The bridge with the lower GN betweenness occupied the position 76. Moreover, the 60% of the bridges occupied the last 18 positions. Now it is clear, why we can not use this measure to discover the network structure: to remove all the bridges we have to remove all the edges. In other words, one can find bridges with high and low betweenness values. Only the recalculation trick can help us in this situation.

The DB arrays look much better. Take for example, the DB with $c = 2$, where the last bridge occupies the position 28. Thus, the problem is solved by removing the 28 edges with the highest differential betweenness values. At this point, we recall that our measure does not require any recalculation.

The DB with $c = 4$ is clearly worse than with $c = 2$ but still much better than the GN case. This is very convenient as far as the computation cost depends on the neighborhood size. The case $c = 1$, which has not been included in the table, was not as bad as we expected. Later we will show that in this case is also possible to discover more information about the structure of the Zachary's graph than with the GN measure.

At this point we have collected some evidences that seem to tell us that the DB is a more robust measure than GN. By this we understand the ability of obtaining a good measurement of the centrality of a bridge in the (common) situation in which there is more than one edge between a given pair of communities. We explain what is happening as follows. Our measure is considering not only the shortest paths, as the GN does, but also the short ones. What do we mean by "short" is controlled by the parameter λ. In other words, the DB does not trade betweenness values among parallel bridges of two communities. We know that the GN betweenness does, and as a consequence it requires the elimination step of the algorithm. Therefore, our new metric gives a better estimation of the actual betweenness value of an edge without been significantly affected by neighboring bridges.

Now we would like to draw the reader attention to an important finding: the differential betweenness can be considered a measure of topological dissimilarity. Indeed, a high value of $DB_{\lambda,c}(i, j)$ means that the neighboors of i are likely to be connected to the neighbors of j through the edge $i \sim j$ (most short paths between pair of vertexes in the c-neighborhood of $i \sim j$ run along this edge). Conversely, a small differential betweenness suggests that the vertexes have many common neighbors. This agrees with a commonsense intuition: similar things have something (the neighbors) in common.

To check out the above claim we computed the dendrograms (with single linkage) of the Zachary's network obtained using the betweenness matrices as dissimilarity metrics. Figure 2 clearly shows that the differential betweenness

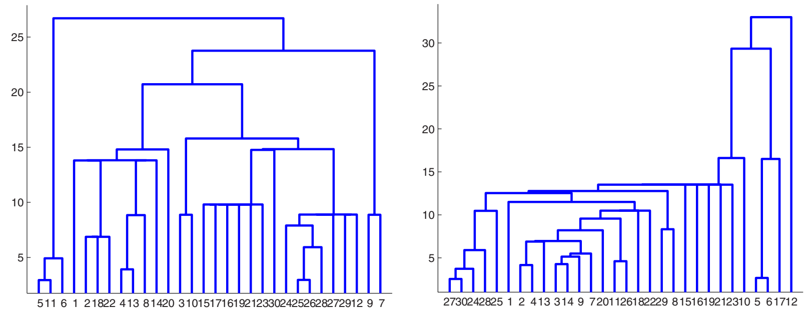

Fig. 2. Dendrograms of the Zachary's network obtained using the betweenness matrices as dissimilarity metrics. Left) Differential with $c = 1, \lambda = 0.01$. Right) Shortest-path. The cophenetic correlation coefficients are 0.44 and 0.32 respectively.

captures much more of the natural clustering of the problem than the GN does. We also computed the cophenetic correlation coefficients, which were 0.44 and 0.32 respectively.

Some readers might think that only one example is not enough to illustrate the actual behaviour of the differential betweenness. Although we completely agree with this way of thinking, we decided to enforce the clarity of the exposition. The presented example is a well known benchmark that uncovers the nice properties of the DB in simple manner. We recall that the aim of the paper is just to point out that there exist networks for which the computation of the DB is a reasonable alternative to existing betweenness centrality methods. Nevertheless, we ensure the readers that it is not difficult to find out many other examples where the DB behaves well.

5 A Note on Time Complexity

There are many interesting issues regarding the computation of the differential betweenness. We will present a detailed discussion in a forthcoming larger paper. Nevertheless, here we want to highlight the importance of the elimination of the recalculation step.

Both the DB and the GN betweenness compute the shortest path matrix of the given network. For unweighted networks this can be accomplished in $O(mn)$ time by breath first search. The GN algorithm repeats this procedure m times due to the recalculation, which gives $O(m^2n)$. It turns out, that the DB matrix can be computed with a worst-case time of $O(mn^2)$ using the computed shortest path matrix. Therefore, the overall complexity is $O(mn) + O(mn^2) = O(mn^2)$. It can be easily seen that for dense graphs we gain one order of complexity. However, for sparse graphs both algorithms give $O(n^3)$. The point here is that the equality is misleading because the constants are quite different. The computation of the DB involves m iterations of $O(n^2)$ simple arithmetic operations with the shortest path matrix, whereas the GN needs lot of graph traversals in each iteration.

We can achieve $O(mn)$ time with the DB if we bound the size of the c-neighborhood, $\mathsf{V}_{c,i-j}$. Assuming that it has a maximum of K nodes, we obtain $O(mn) + O(mK^2) = O(mn)$ or $O(n^2)$ for sparse graphs. Note that in this case, the geodesics are computed in the whole network, but the DB of each edge is estimated in its neighborhood. Alternatively, we might also compute the geodesics inside the neighborhood. This can be done in $O(K^3)$ and $O(K^2)$ time for dense and sparse neighborhoods, respectively. Therefore, under certain special conditions we can even achieve linear time complexity: $O(m(K^3 + K^2)) = O(m)$ or $O(n)$ for sparse graphs. Fortunately, the conditions are more natural than special because many real-world networks has the so-called small-world property, which is in fact all what is needed.

6 Conclusions

We have presented a new result in the general area of clustering and particularly for the special case of finding structures in networks. It consists in a new metric for evaluating the degree of betweenness an edge has in a network. Before our work, it was believed that the design of betweenness measures was not a crucial issue for the development of the field. We can explain this in several ways. On one hand, the poor quality of the shortest paths betweenness metrics made necessary the use of additional methods (as the costly recalculation step) that, indeed, substantially improved the performance of the existing algorithms. Therefore, many researchers and practitioners resigned themselves to accept the resulting high computational cost. On the other hand, the necessary use of a cluster validity index to choose the best cluster decomposition out of the hierarchical tree, shifted the attention of the leading researchers to the use of several powerful optimization methods.

We believe that we need to rethink the whole history. There are a lot of local information that can be extracted from the networks if we are able to find out what are the right features to look at. The use of optimization is of course convenient, in some cases mandatory, but at this point we are confident that a smart combination of the two approaches is the right avenue to follow.

Our differential betweenness metric has several remarkable properties that distinguishes itself as a very promising building block of future community detection algorithms: 1) It is suitable for both weighted and un-weighted networks. 2) It does not need a recalculation step. 3) It is more robust and less sensitive to noise than the existing betweenness metrics. 4) It is a metric of topological dissimilarity.

References

1. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 167–256 (2003)
2. Newman, M.E.J.: Detecting community structure in networks. The European physical journal B, Condensed matter physics 38(2), 321–330 (2004)
3. Newman, M.E.J.: A measure of betweenness centrality based on random walks. Social Networks 27, 39–54 (2005)
4. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Physical Review E 74(036104) (2006)
5. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(026113) (2004)
6. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Physical Review E 69(066133) (2004)