Transitivity Based Community Analysis and Detection

Mohammad Aghagolzadeh, Student Member, IEEE, Hayder Radha, Fellow, IEEE
Department of Electrical and Computer Engineering
Michigan State University, East Lansing, MI, 48823 USA
e-mails: {aghagol1,radha}@msu.edu

Abstract—This paper extends our previous effort in employing transitivity attributes of graphs for social network analysis. Specifically, here we focus on the problem of network community detection. We propose spectral analysis of the transitivity gradient matrix and compare our framework to the modularity based community detection that attracted many network researchers’ attention recently. Previously, we showed that the transitivity attributes of social networks can be analyzed within the resolution of individual links, helping analysts in bridging concepts from the micro- and macro- levels of social network analysis. In this paper, we show that for the problem of network community detection, a key advantage of using transitivity is that it quantifies the degree of community structure independent of the number and sizes of clusters or communities within the network. We employ a Gaussian mixture model in the spectral domain of the proposed transitivity gradient matrix for modeling communities. Performance of the proposed method is compared to the state-of-the-art modularity based community detection over randomly generated networks with social network characteristics such as scale-free degree distributions and high clustering coefficients.

Index Terms—Transitivity matrix, spectral clustering, social networks, community analysis

I. INTRODUCTION

Study of network structures is ultimately reliant on the diversity of tools for quantifying graphical properties. Specifically, for detection of hidden communities in social networks, we intend to focus on triadic (between every three nodes) transitivity attributes of networks rather than their dyadic adjacency attributes that has been the foundation for detection of communities [1], [2]. Communities are often modeled as cliques, corresponding to dense graphical structures with maximum connectivity. However, in sparse networks, intra-community connectivity can diminish, making the detection of communities even more difficult. During the past years, there have been extensive efforts in studying community structure of social networks [3]. In these studies, high transitivity — sometimes referred to high clustering coefficient — has been known as one of the main attributes of social communities. Clearly, using transitivity for finding communities requires a new kernel for spectral graph clustering which is the subject of this paper.

We have shown in [4] that the transitivity attribute of social networks can be analyzed within the resolution of individual links — microscopic entities, — measuring their contribution on the global transitivity index — a macroscopic feature. The basic definition of transitivity was extended to the transitivity matrix encoding impacts of links on the transitivity index of a network (with the sum of impacts equal to the original transitivity index). In this paper, we present spectral analysis of the transitivity gradient matrix that reveals the boundaries of communities and gives an insight into the roles of actors in social networks. The main idea is to extract hidden network structures with high transitivity indices. We evaluate our method over graphs that were simulated using scale-free graph models capable of preserving the natural clustering properties of social networks [5]. Using graphical models provides the ground-truth knowledge of communities, and consequently allows us to compare our method to other methods such as modularity based community detection [1].

This paper is organized as follows. In Section II, we briefly review the new transitivity-framework for analysis of social graphs based on the work developed in [4]. In Section III, we review the utility of spectral analysis for community detection, specifically when triadic transitivity features are employed in spectral analysis of social networks. Simulation results including the comparisons with modularity based community detection are presented in Section IV. Finally, we conclude this paper in Section V.

II. TRANSITIVITY BASED NETWORK ANALYSIS

Transitivity index of a graph is the ratio of the number of triangles to the number of two-stars; a real number between zero and one. Having a high index of transitivity is a prominent attribute of social networks and the existence or lack of transitivity in network substructures has significant implications as noted in pivotal social science works [6]. The transitivity gradient, introduced in [4], captures the amount by which weights of individual links impact the global transitivity index of graphs. We also showed that these gradient attributes could be used to uncover roles of individual links in the binding communities (cohesiveness) or in bridging different communities [4]. It is straightforward to capture these concepts using a differential analysis of the transitivity function, a closed-form extension of the global transitivity index in the form of a continuous function of the adjacency matrix $A$:

$$\tau = \frac{\text{number of triangles}}{\text{number of two-stars}} = \frac{\text{trace}(A^2 A)}{\text{trace}(A^2 K)} \tag{1}$$

$K$ denotes the adjacency matrix of a complete graph of the same size (an all one matrix with zero on its main diagonal).
For a simple illustration, in Fig. 1, we have depicted three binary graphs with their transitivity indices increasing in order from left to right.

![Fig. 1](image)

Transitivity is maximum (equal to one) for complete graphs. Yet, having a unit transitivity index does not imply that a graph is complete. In fact, it is easy to see that the transitivity index of a set of cliques is equal to one as long as cliques do not overlap. One should note that there are no constraints on the number or sizes of cliques which becomes crucial as we discuss later. Cliques are frequently used in social science for modeling communities and a disjoint set of cliques is the canonical form of network analysis. Such canonical form of analysis serves as an idealistic model of network communities. Real social networks are, however, modeled by introducing a level of randomness into the model. Specifically, real networks are usually sparse with bridges connecting different communities. The true value of the transitivity index becomes evident when it is employed as a measure for how close a network is to the described canonical form.

To employ transitivity for detection of communities and analysis of large networks, we need to break down the transitivity index and obtain a high resolution map of transitivity over the entire network. Transitivity index of a link can be [4] defined as the amount of variation in the global transitivity index if that link were to be removed while the rest of the graph was kept fixed. Clearly, removal of links from closed communities would result in reduction of the overall transitivity index while removal of bridges increases the overall transitivity. This complies with the notions from previous paragraphs where we said that the transitivity index of a graph measures how close that graph is to the canonical form.

In mathematical terms, transitivity score of a link is defined as the partial difference of the transitivity function with respect to weight of that link [4]. The derivative of the transitivity function \( \tau(A) \) with respect to \( A \) provides a closed-form derivation of all link transitivity scores:

\[
\frac{\partial \tau}{\partial A} = \frac{3 \text{trace}(A^2K)A^2 - \text{trace}(A^2A)[AK + KA]}{[\text{trace}(A^2K)]^2}
\]  

(2)

We call \( \frac{\partial \tau}{\partial A} \) the transitivity gradient matrix. Note that \( \frac{\partial \tau}{\partial A} \) contains both positive and negative numbers and can be non-zero for links that are not present. We also note that, as discussed in [4], \( \tau \) and \( \frac{\partial \tau}{\partial A} \) are applicable for weighted as well as unweighted graphs. Nevertheless, in this paper we consider binary graphs mainly to simplify the analysis. The transitivity matrix of a graph is defined as the element-wise product of \( \frac{\partial \tau}{\partial A} \) and \( A \): \( T = \frac{\partial \tau}{\partial A} \odot A \) (although, as we discuss later, \( \frac{\partial \tau}{\partial A} \) is used for community detection in this work). An important property of \( T \) is that the sum of its entries is equal to the global transitivity of the graph \( \tau \) [4]. This represents a major departure from the traditional graph transitivity index, which provides a coarse measure for how transitive the overall network is.

III. Transitivity based spectral analysis

In general, spectral clustering algorithms exploit spectral features of any desired similarity matrix to partition data belonging to different clusters (communities for social networks). It is known that for graphs, the similarity matrix can be selected to be the adjacency (link weights) matrix or the Laplacian matrix. In a series of recent publications, surveyed in [1], Newman argues that spectral clustering using the modularity matrix is well suited for the task of community detection in social networks where community sizes could be unbalanced. The modularity matrix is tailored for detecting “modules” in social and biological networks by subtracting the effect of randomness from the adjacency matrix\(^1\).

To summarize the spectral clustering approach to community detection, assume \( \Phi_{n \times n} \) denotes a similarity characteristic matrix of the graph with \( n \) nodes. Then \( p \) principal eigenvectors of \( \Phi \) (associated with \( p \) largest eigenvalues) form a \( p \)-dimensional space where each node is represented by a vector of corresponding elements from the eigenvector matrix. Clustering algorithms such as K-means are applied to these feature vectors to partition the nodes into sets that are believed to represent communities. It is mentioned in [1] that \( p \) principal eigenvectors of \( \Phi \) must be utilized for partitioning the graph into \( p + 1 \) communities. Albeit the low-complexity and the success of K-means for many clustering applications, spectral feature vectors usually form non-spherical clusters which makes K-means unsuitable. As a result, we employ an Expectation-Maximization (EM) framework with the more flexible Gaussian Mixture Models (GMM) for regressing the feature data. More discussion on this matter is postponed to the next section.

In the following subsection, we provide our rationale for using the transitivity matrix for community detection. Noting that spectral clustering is an extension of graph cuts [2], we describe an efficient way to compute graph cuts that result in two partitions with maximum transitivity indices. In other words, we explore a relaxation to the combinatorial problem of finding the smallest graph cut that maximizes the ratio of triangles over the total number of connected triads.

A. Transitivity maximizing cut and spectral clustering

The transitivity score of a link is the amount by which transitivity of the whole graph decreases if that link is removed. However, if multiple links are removed, the sum of their transitivity scores only gives an approximation to the true change in the transitivity index of the resultant graph. This approximation is very close to the true value when the cut size is relatively small compared to the graph size. Also, since

\(^1\)Modularity matrix \( B = A - P \) where \( P_{ij} = d_i d_j / (2m) \) with \( d_i \) denoting the degree of node \( i \) and \( m \) the size of the graph.
computing the exact value of the resultant transitivity index is not the goal of this work, the transitivity matrix provides an efficient way to find the order in which links must be removed.

The approximate transitivity of the resultant graph after a graph cut corresponding to a partition denoted by $\alpha$ is:

$$\hat{\tau}_{res} = \frac{1}{2} \sum_i \sum_j (\alpha_i \alpha_j + 1) T_{ij} = \frac{1}{2} (\alpha^T T \alpha + \tau)$$

(3)

Note that the value of $\frac{1}{2} (\alpha_i \alpha_j + 1)$ is 1 for intra-cluster links and 0 for inter-cluster links. Since we want to maximize the transitivity of the resulting graph, the following optimization problem must be solved:

$$\alpha^* = \arg \max_{\alpha} \alpha^T T \alpha$$

(4)

The new graph cut finds the set of bridges (links between communities) that if removed increases the overall transitivity of the remaining graph. This happens naturally since bridges tend to have negative transitivity scores and are first to be removed. Similar to the ratio cut problem [1], [2], the NP-hard problem of (4) can be approximated by relaxing the integer constraint of $\alpha$. The relaxed solution clearly corresponds to the principal eigenvector of $T$ and partition labels are extracted by a bilevel quantization of the real valued $\alpha$. Multiple partitions can be extracted from $p$ principal eigenvectors of $T$ by performing a vector quantization (in the $p$-dimensional space) using proper clustering algorithms.

Unfortunately, direct application of (4) leads to partitions that are themselves composed of multiple components. The reason is that the transitivity matrix is an overlay of the transitivity gradient matrix $\hat{\tau}$ and the adjacency matrix $A$. Thus, for instance, links with small transitivity scores can be either low weighted links or they can be strong links with neutral transitivity attributes. As a workaround, to resolve the masking effect of the adjacency matrix, we directly analyze the transitivity gradient scores of links:

$$\alpha^* = \arg \max_{\alpha} \alpha^T \hat{\tau} \alpha$$

(5)

We justify the new optimization problem as the following. Let $A$ be decomposed into two (binary or weighted, depending on $A$) components: randomness $R$ and community structure $S$. The idea, similar to the method of modular community detection [1], is that applying spectral clustering over the structure component results in more accurate separation of communities. Based on our discussion regarding the relationship between the transitivity index and the canonical community structure, we expect $S$ to have high transitivity (more than expected for a random graph) and hence the residual $R = A - S$ would have a small transitivity index —assuming $R$ only takes non-negative values. Hence, community structure could be estimated by finding $S$ that minimizes the transitivity of the residual. However, we are using a linear approximation to the transitivity function and linear approximation is only valid within small distances from the initial point. Therefore, the magnitude of change ($-S$) must be infinitely small for the approximation to work. Clearly, $S$ must be in the direction of the gradient $\hat{\tau}$ (ascent) at the current point to minimize the transitivity index at the highest rate. Consequently, spectral analysis of the gradient matrix, as depicted in (5), uncovers the corresponding community structure.

IV. SIMULATIONS

Evaluations are carried over simulated networks with known community structures. The proposed random network simulation is composed of two stages. At the first stage, communities are formed independently with scale-free degree distributions and high clustering coefficients which are both primary properties of social networks [3]. At the second stage, bridges are randomly added to the network such that the number of outgoing links from each community is proportional to its size. Employing random network generators is advantageous for testing community detection algorithms by producing a large number of networks at arbitrary scales with known ground-truth about communities.

Here, we are interested in comparing the performances of spectral-based community detection using $a)$ the modularity matrix [1] and $b)$ the transitivity gradient matrix proposed in this paper. In the following, we present a more detailed description of our random network generator and the spectral-based community detection algorithm. Comparison results are provided later in this section.

A. The random network generator

For generating communities, we employ an extension of the Barabasi-Albert (BA) [7] model for growing scale-free networks proposed by Holme and Kim (HK) [5]. In the original BA algorithm the generated scale-free networks lack transitivity and result in loosely clustered networks. Meanwhile, in the HK model, in order to incorporate high clustering attribute of social networks, an additional step is added to the BA’s preferential attachment step known as the triad formation step. The details of this algorithm is presented in [5] and it is sufficient to note that the HK algorithm can be tuned to satisfy both scale-free and clustering properties of natural networks. We tuned the HK parameters $m = 3$ and $p = 1$ as they appear in the original paper [5].

The network generator first starts off by generating $K$ graphs of arbitrary sizes according to the HK model, representing different communities. Clearly, complexity of community detection —after random bridges are added in the next step— increases as the community size variance gets larger. As noted in [1], community sizes in real social networks could vary in a wide range and Laplacian-based clustering algorithms that try to balance the sizes of clusters (communities) would fail to correctly detect the underlying structure. Here also, we are interested in testing our community detection algorithm under the more realistic unbalanced scenarios. In our simulations, we select $K = 5$ and community sizes are uniformly distributed.

$^2\alpha = \{1, -1\}^n$ is the vector of labels assigning each node to one of the two partitions.
in the range \([5, 50] \in \mathbb{Z}\) representing differences of an order of magnitude.

Once the internal structures of communities are formed, random links are added between different communities. First, we select a community to be active. For \(\kappa\) times, where \(\kappa\) is proportional to the size of community, a random real number \(r\) is uniformly withdrawn from the range \([0, 1]\) and compared to a prespecified threshold \(w\). If \(r < w\), a link is added between a member of the active community to a random node from another community. This process is repeated by selecting another community as active until all communities are processed once. We select \(\kappa = n(n_k/30)\) and \(w = 0.3\) where \(n_k\) is the size of the \(k\)’th community.

\[ B. \text{ Clustering methodology} \]

Finding a two-way partition of a network, also known as a graph cut, through spectral analysis of the characteristic matrix can be done through eigenvector thresholding as explained in [1]. However, capturing multiple communities can be an intricate task, requiring multidimensional modeling of communities in the eigenvector domain, where each node is represented by its corresponding entries of the \(p = K - 1\) most significant eigenvectors. A useful practice is to use Gaussian Mixture Models (GMM) to model multi-modal distributions. Here, each mode corresponds to a community where the node is drawn from. Also, studying the eigen-space of the transitivity matrix, as well as the modularity matrix [1], we found that communities are better captured by ellipsoids than symmetric spheres as in K-means clustering. A fast algorithm for fitting a mixture of Gaussian functions is through the Expectation-Maximization algorithm reviewed in [8]. Although we assume the knowledge of the number of communities in this work, it is possible to estimate this number based on the transitivity matrix [4] which is an ongoing research. For a fair comparison, initial locations of the cluster centers (means of the Gaussian functions) were fixed for the following simulations.

\[ C. \text{ Results} \]

Evaluating clustering results with multiple clusters is itself a challenging problem. In this paper, we resort to a scalar measure of clustering accuracy, the measure of Correct Classification Rate (CCR):

\[
\text{CCR} = \frac{\text{number of correctly classified nodes}}{\text{total number of nodes}} \quad (6)
\]

The two CCR’s are plotted for several generated networks using the model discussed in section IV-A. Fig. 2 shows this plot for 40 sample networks with the mentioned parameters.

In Fig. 2, points below the identity line represent cases for which the utility of the transitivity gradient matrix outperforms the modularity matrix in this problem. The average ratio of CCR for this plot are 82.00% for the transitivity based community detection and 78.57% for the modularity based detection. From the figure, the number of CCR points that are below the identity line are 21 versus 14 points above the line. This represents a 3/2 ratio where transitivity based community detection outperforms modularity based detection. Beside of providing a viable and robust alternative for community detection, transitivity-based analysis of social networks has arguably other key advantages relative modularity-based community detection. In particular, in addition to achieving robust detection of multiple communities, the proposed transitivity matrix provides both microscopic and macroscopic insights about the roles of individual links within a complex social network structure consisting of multiple communities [1]. Such insight and analysis, in particular at the microscopic level, are not readily available from modularity-based detection. Furthermore, under many realistic scenarios, transitivity-based community detection can be achieved without prior knowledge of the number of communities within the overall social network. This could represent a major advantage for transitivity over modularity-based detection.

V. CONCLUSION

Study of social networks has revealed the existence of transitive correlations among triads. In this paper, we propose to use transitivity attributes of graph, unlike density based clustering methods, for community detection. Simulation results demonstrate the robust performance of community detection using spectral analysis of the transitivity gradient matrix, and it is shown to be comparable and sometimes outperforming the state-of-the-art in this field, namely the modularity based community detection.

REFERENCES