ABSTRACT

There has been growing evidence recently for the view that social networks can be divided into a well connected core, and a sparse periphery. This paper describes how such a global description can be obtained from local “dominance” relationships between vertices, to naturally yield a distributed algorithm for such a decomposition. It is shown that the resulting core describes the global structure of the network, while also preserving shortest paths, and displaying “expander-like” properties. Moreover, the periphery obtained from this decomposition consists of a large number of connected components, which can be used to identify communities in the network. These are used for a ‘divide-and-conquer’ strategy for community detection, where the peripheral components are obtained as a pre-processing step to identify the small sets most likely to contain communities. The method is illustrated using a real world network (DBLP co-authorship network), with ground-truth communities.

Index Terms— Social networks, core-periphery structure, community detection, homology, local-to-global

1. INTRODUCTION

The primary assumption underlying all social network analysis [1] is that the behavior of people in a society, is reflected in the properties of combinatorial objects such as graphs and simplicial complexes constructed from observing relationships amongst the people. Perhaps the two most dominant problems have been detecting communities, and analyzing propagation of information flow in the networks. Detecting communities is helpful for identifying organizational structures or similar subgroups of nodes in networks, and there have been several methodologies for community detection proposed in the literature. See [2, 3] for recent surveys of methods, and [4] for comparison of their performance using ground truth information. The study of information flow in networks has been effectively used in epidemiology [5] and belief propagation [6]. Node-wise statistics such as vertex degree, clustering coefficient, or various centrality measures can also be very informative [7, 8], especially when the distribution of such features are considered over an entire network.

Recent advances in technology is providing researchers with access to large and complex networks, and the resources to effectively analyze them. As a result, new information is being uncovered, including the interesting decomposition of a network into a core-periphery structure [9]. While several different interpretations have been presented [10, 11, 12, 13], a common theme is that the core is well connected (expander like) with relatively higher degree vertices, and the periphery is a sparse network with several components.

This paper describes how such a decomposition can be obtained from local “dominance” relationships between vertices. A vertex \( v \) is said to be dominated by a neighboring vertex \( u \) if all of \( v \)’s other neighbors are also neighbors of \( u \). It was shown previously [14] that removing the dominated vertices does not change the topology of the network. The word topology here is used in a precise way, as quantified by homology groups in algebraic topology [15], a major tool in the emerging field of topological data analysis [16]. Intuitively, topology here refers to the “shape” of a given space, such as number of connected components, holes, voids etc.

The core presented here is what remains after iteratively removing these dominated vertices. It is shown that the shortest distances between all the pairs of vertices in the core are preserved. This conforms with our intuition of a core network, where mutual relationships are independent of what happens outside the core network. Experiments also verify the “expander-like” property of the core, i.e., the network cannot be divided into two equal (or almost equal) parts by removing small number of edges. The periphery consists of many components, which we call peripheral groups. The peripheral groups obtained in this decomposition correspond very closely to communities, both with respect to well-established measures of how ‘community-like’ a group is, as well as when compared to ground-truth communities. We illustrate these results using a coauthorship network from DBLP as an example. Furthermore, such a decomposition of the network into core and periphery may be accomplished distributively [14]. These properties of core and periphery are consistent with other notions of such a decomposition.
presented previously [12, 13].

We will proceed as follows: Section 2 outlines the distributed, iterative algorithm for collapsing a network on the basis of node dominance. In Section 3 the notion of a core-periphery decomposition of a network is described, and it is shown that the output of the node dominance algorithm displays properties expected from a core. Section 4 then shows how the connected components in the periphery can be interpreted as communities in the network, with empirical results verifying this interpretation.

2. COLLAPSING A NETWORK USING NODE DOMINANCE

The algorithm we employ was originally developed ([17] [18]) as a homology-preserving collapse of a simplicial complex, and was performed as a pre-processing step to reduce computational complexity for purposes of topological data analysis [19].

2.1. Simplicial complex representation

A simplicial complex is a higher-dimensional analogue to a graph, and consists of vertices and edges (pairs of vertices), as well as sets of n-tuples of vertices, such as triangles, tetrahedra, etc. A set of k + 1 vertices is called a k-simplex (plural: simplices), and a simplicial complex K is a set of simplices with the additional property that for any simplex σ ∈ K if τ ⊆ σ is a subset of σ (i.e. τ is a simplex defined by a set of vertices which is a subset of the vertices that define σ), then τ ∈ K. Like a graph, a simplicial complex lends itself well to discrete combinatorial representation, and computations on it may be performed in terms of matrix operations. The homology of a simplicial complex K is a sequence of vector spaces (H₀, H₁, ..., Hk), where the rank of the k-th vector space counts the number of k-dimensional ‘holes’ in the complex. A hole can be thought of as an empty space that is surrounded by a chain of (k − 1)-simplices. For example, if a set of edges forms a cycle but is completely ‘filled-in’ by triangles, then the cycle is homologically trivial and doesn’t correspond to a hole. One may also think of such filled-in cycles as being collapsible to a point.

Given a graph G(V,E), one way to build a simplicial complex from it is to take the flag complex, which includes a k-simplex \{v₀, v₁, ..., vₖ\}, if (vᵢ, vⱼ) ∈ E for all i, j ∈ \{0, ..., k\}, i.e., whenever all possible pairs of vertices in the k-simplex are edges in the graph G(V,E). Given such a flag complex, we interpret its nontrivial homology as the essential structure of the network, while the portions which are trivial/collapsible do not contribute to overall structure, but represent locally well-connected groups that are easily separable from the ‘core’ structure of the network (i.e. communities).

2.2. Node dominance

For a node v in the graph G(V,E), its neighbor set

\[ N(v) = v \cup \{ u \in V : (u, v) \in E \}, \]

consists of all nodes attached to v by an edge, as well as v itself. A node v is said to be dominated by one of its neighbors w if \( N(v) \subseteq N(w) \). The neighbor set of v is contained in the neighbor set of w. Removing a node that is dominated does not change the homology of the complex (i.e. all holes are preserved). This was the primary motivation in [14] for using node dominance to simplify the network.

The focus in this paper however, is the graph-theoretic properties of the core and periphery obtained by iteratively removing dominated nodes until there are no more dominated nodes present (noting that a node may become dominated at some point only after other nodes have been removed from the network). The nodes remaining are designated as ‘core’, and the nodes that had been removed are designated as ‘periphery’. See [14] for full details of the distributed algorithm.

3. PROPERTIES OF CORE AND PERIPHERY

This section describes the properties of core and periphery obtained by the decomposition described in the previous section, along with empirical verification of these properties.

3.1. Dataset

As a running example through this paper, a large DBLP coauthorship network (available from the Stanford SNAP database

| Nodes in core: | 71,018 |
| Nodes in periphery: | 246,062 |
| Nodes (total): | 317,080 |
| Edges within core: | 318,741 |
| Edges within periphery: | 274,367 |
| Edges between core and periphery: | 456,758 |
| Edges (total): | 1,049,866 |
| Mean degree: |
| Entire network: | 6.62 |
| Core (w.r.t entire network): | 15.41 |
| Core (w.r.t. core): | 8.98 |
| Periphery (w.r.t entire network): | 4.09 |
| Periphery (w.r.t periphery): | 2.23 |

Table 1. Descriptive statistics for the DBLP coauthorship dataset, and its core-periphery decomposition.
[20]) is used. Summary statistics about the full network, as well as its core-periphery decomposition are given in Table 2.2. This network also has information about its ground-truth communities, where communities are defined as connected components of authors within the same publishing venue, as proxies for scientific communities.

3.2. Properties of the core

The original graph/network is denoted as $G = G(V,E)$, and the vertex set $V$ is partitioned into the core $V_C$ and periphery $V_P$. The graphs induced by $V_C$ and $V_P$ are denoted by $G_C$, and $G_P$ respectively.

3.2.1. Shortest paths are preserved

**Theorem 3.1** For two nodes $v_1, v_2 \in V_C$, let $d_{G}(v_1, v_2)$ and $d_{G_C}(v_1, v_2)$ denote the corresponding shortest path distance (in hop length) in $G$ and $G_C$ respectively. Then, $d_{G}(v_1, v_2) = d_{G_C}(v_1, v_2)$.

**Proof** For any graph $G'$, let $v_j$ be dominated by its neighbor $v_i$. Consider any shortest path $p = \ldots, v_k, v_j, v_i, \ldots$, passing through $v_j$. Note that $k, l \neq i$ [Proof by contradiction: $p = \ldots, v_i, v_j, v_i, \ldots$ could be replaced by shorter path $\ldots, v_i, v_l, v_j, v_i, \ldots$ because $v_l \in N(v_j)$, since $v_l \in N(v_i)$ and $v_i$ dominates $v_j$. So $p = \ldots, v_k, v_j, v_i, \ldots$ can be replaced by the path $p' = \ldots, v_k, v_i, v_i, \ldots$, which is the same length as $p$ but does not contain $v_j$. Therefore, all the shortest path lengths in $G'$, where $v_j$ is not the source or destination, are preserved when $v_j$ is removed.

3.2.2. The core is expander-like

Two common measures of how expander-like a graph is, are edge expansion and vertex expansion [21], but obtaining these quantities is computationally prohibitive for large graphs. A third commonly-accepted measure of a graph’s expansion is the spectral gap (the magnitude of the smallest nonzero eigenvalue) of the normalized graph Laplacian

$$\mathcal{L} = I - D^{-1/2}AD^{-1/2},$$

where $A$ is the adjacency matrix for the graph, and $D^{-1/2}$ is a diagonal matrix with entries $1/\sqrt{d_i}$, where $d_i$ is the degree of the $i$-th node. A larger spectral gap indicates a more expander-like graph [22].

On the DBLP coauthorship network, the spectral gap of the entire network is $\lambda_2(G) = 0.0027$, while the spectral gap of the core is $\lambda_2(G_C) = 0.0432$ (about 16 times larger).

Since it is difficult to interpret the magnitude of the spectral gap when comparing networks of different sizes, 100 random subgraphs of the same size as the core were obtained from the entire network, using the forest fire subsampling method [23] with forward burning probability $p_f = 0.6$. The spectral gap for each of these subgraphs was computed, and over the 100 simulations, the average spectral gap was 0.0102 with standard deviation 0.0037. This means that the core is significantly more expander-like than a random sampled subgraph of the network would be.

3.3. Measures of community quality

The connected components in the periphery, which are referred to here as peripheral groups, score well with measures associated to communities. The two such measures considered in this paper are described here. A very intuitive measure of community quality is conductance [2]. For a set of nodes $S$, the conductance is measured by the ratio of the number of edges leaving the set to the total number of edges associated to communities. The two such measures considered in this paper are described here. A very intuitive measure of community quality is conductance [2]. For a set of nodes $S$, the conductance is measured by the ratio of the number of edges leaving the set to the total number of edges associated with the set:

$$\text{cond}(S) = \frac{c_S}{\sum_{v \in S} \deg(v)}$$

where $c_s = |\{(u, v) : u \notin S, v \in S\}|$. Another function measuring community quality is one that focuses more on the internal connectivity of a set: triangle participation ratio (TPR)

$$\text{TPR}(S) = \frac{|\{(u \in S : v, w \in S, (u, v), (u, w), (v, w) \in E)\}|}{|S|}$$

is the fraction of nodes in the set that belong to a triangle. Both conductance and triangle participation ratio have been shown to have good community-detection performance [4].

3.4. Properties of the periphery

The distribution of conductance and TPR for the peripheral groups (of size $\geq 6$) from the DBLP data set are shown as the solid lines in Figure 1. We can see that the peripheral groups display relatively good (low) conductance values, and so are well-separated from the rest of the network. Many of the peripheral groups display good (high) TPR scores, but a number of others are only moderate. This is likely due to the node dominance only requiring that the peripheral group be collapsible onto the core through a sequence of node dominance collapses, so not all peripheral groups are strongly internally connected. For example, some display longer ‘tails’ or star-like sections which do not include any triangles.

4. COMMUNITY DETECTION

Typically, an objective function (such as conductance, described in the above section) is chosen to describe how ‘community-like’ a group of nodes is, and an algorithm is developed to find a partition of the network into communities which score well with respect to this function (finding the absolutely optimal partition is usually not computationally feasible).
Fig. 1. Distribution of conductance (left) and triangle-participation-ratio (right) for the peripheral components containing at least 6 nodes (solid line), for the 5000 best quality ground-truth communities (dotted line), and for all 13,477 ground-truth communities (dashed line).

Fig. 2. Examples of some peripheral components containing ground-truth communities. The nodes in the peripheral component and internal edges are drawn in black, and the neighboring nodes in the core along with the edges connecting them to the peripheral component are drawn in grey. Nodes in the ground-truth community contained in the component are highlighted by boxes.

The DBLP coauthorship dataset also contains information about ground-truth communities in the network. The communities are defined as connected components of authors within the same publication venue (which are meant to act as proxies for scientific communities). There are 13,477 ground-truth communities provided in total, but 5000 of these communities are labeled as being of the highest quality (see [4] for complete details). The ground-truth communities tend to be very well internally connected (high TPR), but are not always well-separable from the rest of the network (can have high conductance). The conductance and TPR are shown in Figure 1 for all the ground-truth communities (dashed line) and the 5000 highest-quality ground-truth communities (dotted line).

The core-periphery decomposition actually gives quite valuable information about where the ground-truth communities reside in the network. Out of the 5000 highest-quality communities, only 6 (0.01%) are contained entirely in the core, but 3251 (65.0%) of them intersect exactly one peripheral component, and 4204 (84.0%) of them intersect at most two peripheral components. This indicates that the core-periphery decomposition can be used as a pre-processing step to identify small candidate sets most likely to contain meaningful communities. If the candidate sets are defined as the peripheral components plus their neighboring nodes in the core, then 3124 (62.5%) of the ground-truth communities are contained in exactly one candidate set. Embeddings of a number of peripheral groups are plotted in Figure 2, with the nodes belonging to the ground-truth communities contained in them highlighted by boxes. The embeddings were obtained by using multidimensional scaling on the graph distance between nodes, plus small random errors to avoid pairs of distances being exactly equal. Further, the distances between nodes in the peripheral group and those in the core were increased for purposes of visualization.

5. CONCLUSION

This paper describes an iterative procedure using node dominance to collapse a network onto its core, thus decomposing it into core and periphery. The core obtained from this decomposition corresponds well with existing ideas about the ‘core’ structure of a network: the nodes have high degree, but relatively few triangles are present, which gives the core an expander-like quality (as further evidenced by the spectral gap of the core being significantly larger than the spectral gap for other equally-sized subsamples of the network). The peripheral components of this core-periphery decomposition are very community-like themselves, in terms of conductance, and can be used to identify small subsets of the graph most likely to contain ground-truth communities, thus providing an efficient algorithm for community detection.

Material presented here also provides evidence behind the intuition that non-trivial topological features in (the flag complex of) a network correspond to essential network structure. Thus, performing a homology-preserving collapse yields the core of the network, while sections with trivial homology correspond to communities.

The structure discovered in the core behooves us to investigate its role in information propagation in the network.
Another interesting notion of a core which has not been considered here is its stability in time varying networks. In some cases it may not be desirable to strictly preserve all homology, and the node dominance condition could be relaxed to include partial dominance (which would no longer be guaranteed to preserve homology), or edges could be added between pairs of nodes two hops adjacent. The latter relaxation would allow small non-trivial cycles to be filled in, and could be used to detect a deeper core, nested within the original. These are topics for future exploration.

REFERENCES


